



Reichman University

Efi Arazi School of Computer Science

M.Sc. program - Research Track

CompanyName2Vec: Company Entity Matching Based on Job Ads

by

Ran Ziv

M.Sc. dissertation, submitted in partial fulfillment of the requirements for the M.Sc.
degree, research track, School of Computer Science, Reichman University

November 2021

This work was carried out under the supervision of Dr. Michael Fire from the department of Software and Information Systems Engineering, Faculty of Engineering, Ben Gurion University, and with guidance from Dr. Ilan Gronau from the Efi Arazi School of Computer Science, Reichman University

Abstract

Entity Matching is an essential part of all real-world systems that take in structured and unstructured data coming from different sources. Typically no common key is available for connecting records. Massive data cleaning and integration processes require completion before any data analytics or further processing can be performed. Although record linkage is frequently regarded as a somewhat tedious but necessary step, it reveals valuable insights, supports data visualization, and guides further analytic approaches to the data. Here, we focus on organization entity matching. We introduce `CompanyName2Vec`, a novel algorithm to solve company entity matching (CEM) using a neural network model to learn company name semantics from a job ad corpus, without relying on any information on the matched company besides its name. Based on a real-world ground truth dataset, we show that `CompanyName2Vec` outperforms other evaluated methods with average accuracy at 1 of 0.894.

TABLE OF CONTENTS

List of Tables	vi
List of Figures	vii
Chapter 1: Introduction	1
Chapter 2: Related Work	6
2.1 Entity Matching	6
2.1.1 Supervised Approaches	6
2.1.2 Unsupervised Approaches	7
2.2 Distance Metrics	7
2.2.1 Character-Based Similarity Metrics	8
2.2.2 Token-Based Similarity Metrics	9
2.2.3 Phonetic Similarity Metrics	9
2.3 Document Fingerprinting	10
2.4 Text Embedding	10
Chapter 3: Method	12
3.1 Hiring Company Names Fingerprinting	12
3.2 Company Name Embedding	13

Chapter 4: Experimental Study	16
4.1 Dataset	16
4.2 Evaluation Process and Performance Metrics	17
Chapter 5: Results	19
Chapter 6: Discussion	23
Chapter 7: Conclusions and Future Work	25
References	27
Appendices	32
Appendix A: Additional Results	33
Appendix B: References to project Code and Datasets	35

LIST OF TABLES

2.1	Company Names Synonyms Ratio Similarity vs. Partial Ratio Similarity . .	9
3.1	For example, we provide four fingerprint values and the lists of hiring company names posted with the same job ad's fingerprint	14
5.1	Algorithms' Performance Comparison	20

LIST OF FIGURES

1.1	Flowchart of the implemented methodology.	5
5.1	Fingerprints Distribution by Company Names	20
5.2	Synonym distribution of the following companies as were calculated by the CompanyName2Vec algorithm and reduced into two dimensions with t-SNE: ABM Industries, ACCO Brands, HCA Healthcare, Home Depot, J.B. Hunt Transport Services, Inc., JPMorgan Chase & Co., Lowe’s Inc., and PepsiCo	21
5.3	A plot of PepsiCo synonyms’ vectors after were reduced into two dimensions	21
5.4	A plot of HCA Healthcare synonyms’ vectors after were reduced into two dimensions	22
5.5	Distribution of company names exist in job ad corpus	22
A.1	A plot of Lowe’s Inc. synonyms’ vectors after were reduced into two dimensions	33
A.2	A plot of Home Depot synonyms’ vectors after were reduced into two dimensions	34

CHAPTER 1

INTRODUCTION

Enterprise Business intelligence systems have emerged as a disruptive technology and innovative solution to the global economy [1]. Business intelligence became an emerging and fast-growing field in the past years [2]. Big data and its emerging technologies, including business intelligence and big data analytics systems, have become a mainstream market adopted broadly across industries, organizations, and geographic regions to facilitate data-driven decision making and significantly affect the way that decision-makers, such as CEOs, operate and run their business [2, 3]. One of the fundamental functionalities a business intelligence system must have is the ability to integrate many data sources [2, 4]. The integration of multiple sources usually requires linking between the significant entities that exist across data sources and systems. Usually, those entities function as a “primary key” in each system [4]. The technique used for performing such linkage is commonly referred to as “Record Linkage”, “Data Deduplication”, “Object Matching”, or “Entity Matching” [4, 5].

Entity Matching (EM) is a fundamental task in data integration scenarios. EM is the task to identify semantically equivalent entities referring to the same real-world object (e.g., persons, products, companies) within one data source or between different sources. EM is also a core technique for data cleaning [5, 6] and data integration [7, 8]. Accurate and fast entity matching has huge practical implications in a wide variety of commercial, scientific and security domains [9]. EM solutions can be divided into three separate groups: supervised, unsupervised, and semi-supervised approaches [10, 11]. Supervised approaches require labeled training sets or predefined thresholds on which to base their decisions. Unfortunately, in most real-world cases, the variety of patterns that can be observed are not feasible to be captured in a training set. Therefore, these solutions are quite limited. Unsu-

pervised approaches help find previously unknown patterns in a dataset without pre-existing labels and are based on clustering algorithms that group together items with high similarity [12]. Semi-supervised approaches fall between the unsupervised approach (with no labeled training data) and a supervised approach (with only labeled training data). Semi-supervised approaches are based on a small set of labeled data and a large set of unlabeled data.

Many enterprise business intelligence applications require the integration of multiple data sources. In such applications, a company name is one of the most crucial entity attributes to be linked [4]. A company usually exists in many of the enterprise systems. For example, a company can represent a potential customer, a sales lead in the sales system, an existing customer in the customer relationship management system, a supplier in the logistics system, etc. Although a company name can be noisy, in most cases, this is one of the most potent properties for company entity matching (CEM) due to the lack of unique shared identifiers across datasets [4].

Determine whether a new company name is in a database, and if so, which existing record it refers to is a typical problem business intelligence applications need to solve given records indexed by company names and a new company name. This problem is an instance of entity matching problem. It is a challenging problem because people do not consistently use the official name, but use abbreviations, synonyms, different order of terms, different spelling of terms, short form of terms, and the name can contain typos or spacing issues.

Many of the existing methods use more company properties other than the company name, such as location, primary phone number, industry, website URL, and more to improve CEM results. Such company information is usually costly to acquire on a large scale. The largest public source for such company information is DBpedia [13]. DBpedia contains about 65,000 company entities worldwide, derived from the English version of Wikipedia [4]. There are several companies worldwide that hold a much larger dataset of company information, such as Dun and Bradstreet and Infogroup [14]. These companies do not provide the complete dataset but provide a limited paid service for company

information enrichment, mainly for sales and marketing purposes.

In this study, we developed `CompanyName2Vec` (see Chapter 3), a novel algorithm to solve CEM using a neural network model to learn company name semantics from a job ad corpus. Once trained, such a model can suggest synonymous company names. As the name implies, `CompanyName2Vec` represents each different company name as a vector. The vectors are chosen carefully such that a simple mathematical function, like cosine similarity between vectors, indicates the level of semantic similarity between the company names represented by those vectors. We used `CompanyName2Vec` for CEM. For this purpose, we created a real-world dataset that consists of the largest companies in the US and their synonyms. We used this dataset to evaluate the `CompanyName2Vec` algorithm and compared it with other known entity matching methods.

To develop the `CompanyName2Vec` algorithm, we needed to have a relatively large corpus of company names. Job ads consist of company names and are available publicly on a large scale but are very noisy (see Section 3.1) due to how job ad distribution works. Companies distribute their jobs to multiple job boards and sometimes to recruitment and staffing agencies to reach as many job seekers as possible. In many cases, the process for posting a job ad on job boards requires the employer to manually type the job information, including company name, location, title, description, and more, which causes some inconsistency and duplication of job ads. For example, Tesla’s job ad can be posted with several employer names like `tesla`; `Tesla, Inc.`; `Tesla Motors`; `Tesla Motors, Inc.`; and `Ursus`, which is a staffing and recruitment agency. `CompanyName2Vec` leverages the availability of job ads publicly (see Section 4.1) and their inconsistency to learn company name semantics (see Section 3.2).

The method developed consists the following steps (see Figure 1.1): First, we collected a large dataset of job ads (see Section 4.1). Second, we used a fingerprinting technique and created a list of company names that posted the same job (see Section 3.1). Additionally, we cleaned the data to filter out non-valid job ads and company names to create a ground

truth dataset. Next, we pre-processed this cleaned dataset and created train and test sets (see Section 3.2). Afterward, we trained a neural network model to learn company names semantics and transform company names into a vector representation and used cosine distance for measuring the similarity between different vectors, where two company names that represent the same company will have similar vectors (see Section 3.2). For example, employers write their company name in many ways, sometimes ignoring common suffixes, like Inc., and LLC, which are used to define the company legal entity type but are less important when posting a job ad or tagging the company name in a social media post.

Lastly, we created a real-world dataset of the largest 1,000 companies in the US and their synonyms (see Section 4.1). We used this dataset to evaluate our method's performance by measuring the success rate at k of matching company synonyms with their mapped canonical company name (see Section 4.2).

Our main contributions are twofold:

- CompanyName2Vec, a novel algorithm to create a mathematical representation for a company name that holds the naming semantics, which enables it to measure similarity between any given company names without relying on any information on the matched company besides its name (see Chapter 3).
- An end-to-end, highly scalable, enterprise-grade system that uses CompanyName2Vec algorithm to suggest company name synonyms to solve company entity matching (see Chapter 4).

The remainder of the paper is organized as follows: In Chapter 2, we provide an overview of related studies. Next, in Chapter 3, we present the methods used to create the CompanyName2Vec embedding model. Afterwards, in Chapter 4, we describe the experimental study we conducted to construct the model and evaluate it. Subsequently, in Chapter 5, we present the obtained results. Then, in Chapter 6, we discuss the obtained results. Lastly, in Chapter 7, we present our conclusions from this study and offer future

research directions.

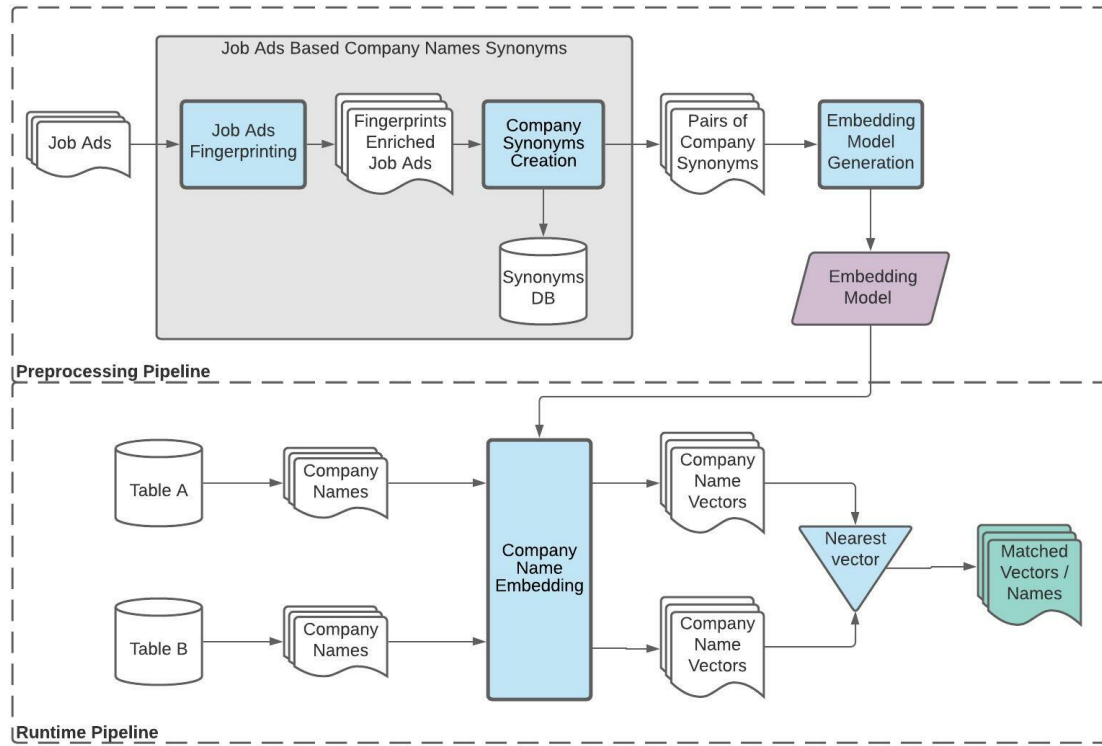


Figure 1.1: Flowchart of the implemented methodology.

CHAPTER 2

RELATED WORK

Entity Matching (EM) is an important task, which many researchers in the past have addressed [5, 15, 16, 17]. In this study, we addressed a particular case of EM, namely, the matching of company names across multiple datasets and systems. In section 2.1, we present the current approaches concerning EM. In section 2.2 we describe several distance metrics methods and techniques. In section 2.3 we describe methods for document fingerprinting. Lastly, in section 2.4, we explain embedding with an emphasis on text embedding.

2.1 Entity Matching

EM addresses matching entities between different data sources or deduplication of entities within a single source. EM uses algorithms to both detect duplicates in data and resolve them. As described above, EM solutions can be divided into three separate groups: supervised, unsupervised, and semi-supervised approaches. In the following subsections, we describe the different methods for EM.

2.1.1 Supervised Approaches

The supervised approaches are based on training data in the form of record pairs, pre-labeled as match or not match. In 1969, Ivan Fellegi and Alan Sunter [18] denote two classes: A class M , which represents matches, and a class U , which represents non-matches. There are three main approaches: Rule-based, learning-based, and distance-based.

Rule-Based. One way of performing entity matching is by setting a set of rules that identify the conditions that would make two records be considered as matched. For instance, matching a customer entity which includes its name and full address, it could be of the form: $(name, edit, =, 1) \wedge (address, edit, >, 0.7)$, which indicates that two customer

records would match if their names are fully matched and their address and city attributes are more than 70% similar.

Learning-Based. Learning-based approaches [5] use training sets that consist of pairs labeled as Match or Non-match. Each pair includes a comparison vector that represents the comparable attributes of the two items in the pair. Assuming the density function for the classes M and U are different, the EM problem can be treated as a Bayesian inference problem.

Distance-Based. Distance-based approaches [5] theoretically do not need labeled data. In these approaches, a distance metric is defined between data items. A decision is made based on whether or not this pair is a match or not based on the distance between two items and a threshold. This threshold can be set by making a reliable estimation. A good threshold can improve the results. Therefore using a training set for setting this threshold might be desirable.

2.1.2 Unsupervised Approaches

The idea behind unsupervised approaches is that similar comparison vectors correspond to the same class [5]. Unsupervised learning for EM has its roots in the probabilistic model proposed in 1969 by Ivan Fellegi, and Alan Sunter [18]. When there is no training data to compute the probability estimates, it is possible to use variations of the Expectation-Maximization algorithm [19] to better identify appropriate clusters in the data.

2.2 Distance Metrics

A common source for mismatches in database entries is the typographical variation of string data. One of the methods to deal with typographical variations is to measure the similarity between two different strings. In this section, we describe string matching techniques that have been applied in EM.

2.2.1 Character-Based Similarity Metrics

The character-based similarity metrics method is designed to manage typographical errors efficiently. A popular character-based method is the Levenshtein distance (also referred to as Edit Distance) [20]. This method measures similarity between two strings by counting the minimum number of operations required to transform one string into the other, including insertion, deletion or substitutions. Damerau-Levenshtein distance [20, 21] is a variation of the Levenshtein distance where a transposition of two characters is also considered to be an elementary edit operation in addition to insertion, deletion, and substitution. Jaro distance [22] is based on the number of common characters and the number of transpositions in two strings. Winkler distance [22] is a variation of the Jaro distance and gives higher scores to strings that share the same prefix.

In terms of company names, Levenshtein Distance seems to work well for a single word or a much longer text, but not for just for a few words [23], as in our use case. For example, “New York Yankees” and “Yankees” are clearly referring to the same company name, but “New York Mets” and “New York Yankees” are clearly referring to different ones. Yet, the score of the “wrong” match is higher than the “right” one. For such cases, there is an heuristic called “best partial” [24], which uses partial matching logic. This logic is implemented in the Fuzzy-Wuzzy package [24], which is based on Levenshtein Distance [20] to calculate the differences between two different strings. Given two strings X and Y , let the shorter string X be of length m . It finds the ratio similarity measure between the shorter string X and every substring of length m in the longer string Y , and returns the maximum of those similarity measures. So in the case of “Yankees” and “New York Yankees” the score will be higher than the score of “New York Mets” and “New York Yankees” (see Table 2.1), since the substring “Yankees” is wholly contained in the string “New York Yankees”.

Table 2.1: Company Names Synonyms Ratio Similarity vs. Partial Ratio Similarity

String X	String Y	Ratio Similarity	Partial Ratio Similarity
YANKEES	NEW YORK YANKEES	61%	100%
NEW YORK METS	NEW YORK YANKEES	76%	69%
NEW YORK METS	NEW YORK MEATS	96%	92%

2.2.2 Token-Based Similarity Metrics

The method of Character-based similarity metrics measures similarity depending only on the appearance and sequence of characters, while token-based similarity metrics first tokenize two strings into sets of tokens (words) and only then compute the similarity between the two sets. Token-based similarity metrics are robust in measuring the similarity of full names, for example, where the order of the first name and last name may change from one string to another. A standard method for Token-based similarity metrics is *TF-IDF* (Term Frequency / Inverted Document Frequency) [25], a numerical statistic method that intends to reflect how important a word is to a document in a collection of documents. *Cosine similarity* [25] measures the similarity of strings by transforming words into vectors, where the frequency of a word is a dimension in the vector. Cosine similarity measures the similarity between two strings by measuring the angle between the vectors. Token-based similarity metrics methods do not take misspellings into account. Therefore those methods are usually used together with character-based methods to determine whether two tokens are similar enough [25].

2.2.3 Phonetic Similarity Metrics

Strings may be phonetically similar even if they are not similar at the character or token level. Different from character-based and token-based similarity metrics, phonetic similarity metrics are limited to a string-based representation. An example of this is the name

Claire. It has two alternatives, Clare and Clair, which are both pronounced the same. Soundex [26, 27], for example, is one of the best known phonetic encoding algorithms for indexing names by sound as pronounced in English. Soundex keeps the first letter and converts the rest of the string into numbers according to a phonetic encoding table. Additional string comparison methods can be found in a thorough survey written by Peter Christen [28].

2.3 Document Fingerprinting

Among digital data, documents are the easiest to copy and remove any signatures or fingerprints embedded, which make the pirating the hardest to detect [29]. Anyone can retype a document or copy a part of it. Document fingerprinting is concerned with accurately identifying and copying, including small partial copies, within large sets of documents [29]. Our study uses fingerprinting to find hiring company synonyms in the job ads dataset by calculating a job ad fingerprint and looking for hiring company names with the same fingerprint. A Checksum, for example, is a small digest of the entire document. This method is simple and sufficient for detecting exact copies. Still in some cases, as well as for our study, there is a requirement in a method that is more local and detects partial copies like the Winnowing algorithm [30], which selects the q -gram whose hash value is the minimum within a sliding window of q -grams. Winnowing is used for text clustering [31], and detecting plagiarism, like the comparison Agung Toto Wibowo published in 2013 [32] which detects plagiarism fraud on Bahasa Indonesia documents, or Moss by Alex Aiken [33] which is a system for measuring software similarity, used by Stanford university for detecting plagiarism in programming classes.

2.4 Text Embedding

In this study, we used embedding to create a mathematical representation for a company name that holds the naming semantics. When some object x is said to be embedded in

another object y , the embedding is given by some injective and structure-preserving map $f : X \rightarrow Y$. We used text embedding, similarly to what Yoshua Bengio proposed in 2001 by [34], in the form of a feed-forward neural network language model. Modern methods use a simpler and more efficient neural architecture to learn word vectors, like word2vec [35, 36], and GloVe [37], based on objective functions that are explicitly designed to produce high-quality vectors. Neural embedding learned by these methods have been applied in a myriad of NLP applications, including initializing neural network models for objective visual recognition [38], or machine translation [39, 40], as well as directly modeling word-to-word relationships [35, 41, 42, 43], Paragraph vectors, or doc2vec, were proposed in 2014 by Quoc Le and Tomas Mikolov [44] as a simple extension to word2vec to extend the learning of embedding from words to word sequences. Doc2vec¹ is agnostic to the granularity of the word sequence, and it can equally be a word n-gram, sentence, paragraph, or document. One of the benefits of using dense and low-dimensional vectors is computational. The main benefit of this dense representation is generalization power. If we believe some features may provide similar clues, it is worthwhile to give a representation that can capture these similarities. We used text embedding to learn company names semantics and to measure similarity.

¹The term doc2vec was popularized by Gensim, a widely-used implementation of paragraph vectors: <https://radimrehurek.com/gensim/>

CHAPTER 3

METHOD

The method described below deals with CEM by creating a generic open solution that matches company names. Unlike many other existing solutions that utilize expensive commercial datasets, the presented method’s novelty uses job-ads data, and it requires no company’s information for matching besides its name.

Our method consists of two main parts: first, the creation of pairs, in which a pair consists of synonyms for the same company based on an analysis of large-scale job ads data (see Section 3.1); second, the algorithm `CompanyName2Vec` for constructing an embedding model that utilizes these pairs of synonyms. The embedding model takes as input a company name and returns as output meaningful vector representation, where two company names that represent the same company will be relatively close vectors in terms of the Cosine distance (see Section 3.2).

3.1 Hiring Company Names Fingerprinting

Given a job ads corpus, we first create and enrich each job ad with a unique job ad identifier (aka job ad fingerprint) using a fingerprinting technique. Second, we use the job ad fingerprint to identify company name synonyms. We group the job ad data by job ad fingerprint and create a list of company names posted with the same fingerprint, or in other words, company names that were published within the same job ad. Then, we remove duplicated synonyms and create a list of pairs, where each pair includes company names that were posted within one or more job ads.

A particular case we need to handle is staffing and recruitment agencies, which often publish their customers’ job ads anonymously and use the staffing or recruitment agency name instead. Ideally, each pair includes company names that refer to the same company.

Therefore, to reduce such noise, we filter out staffing and recruitment agencies’ names using naive text matching. Our analysis found that much of the noise can be reduced by filtering company names which include the following strings: “staff”, “recruit”, “jobs”, or “unknown”.

Given a dataset of job ads, our method’s first step is to create a unique identifier (aka fingerprint) for each job ad. This component calculates a job ad fingerprint based on the job description. We used a local approach to calculate fingerprints [30]. First we calculated a Winoing list of fingerprints. We used the following parameters: $kgram_len = 4$, $window_len = 5$, $base = 10$ and $modulo = 1000$. Then, we utilized the MD5 hashing algorithm [45] to convert the list of fingerprints into a single short 128-bit job identifier to improve computation and debugging efficiency.

3.2 Company Name Embedding

The purpose of collecting company synonyms from job ads is to create a dataset for the method’s embedding model. An outcome of it is the creation of a large database of company synonyms. We utilize the fingerprints to group company names with the same job ad’s fingerprint to construct training and testing sets. After grouping the company names by fingerprint, it transforms the group of names into a set of pairs (i_{name}, j_{name}) where i_{name} ’s and j_{name} ’s Job Ads have at least one Job Ad’s fingerprint in common.

It is important to note that a company may have several fingerprints, at least as many as the number of open positions it tries to fill.

In addition, we removed duplicated pairs from this extensive list of names pairs, which created when two company names have more than just a single job ad fingerprint in common.

In the algorithm’s second step, we utilize an embedding algorithm (see Section 2.4) to capture the semantics of company names by placing semantically similar company names close together in the embedding space. We call this embedding solution Compa-

Table 3.1: For example, we provide four fingerprint values and the lists of hiring company names posted with the same job ad’s fingerprint

Fingerprint	Company Names
9abc97a34a 7c5630b0673083fb9c61c1	99 Cents Only Stores; 99 Cent Only; 99 Cents Only; 99 Cents Only Stores LLC A.O. Smith; AO Smith; AO Smith Corporation; A. O. Smith; A. O. Smith Corporation; A. O. smith; Smith (A.O.) Corporation
c9245756a0edca343c96a1a3b8762fc0	ADP; ADP Automatic Data Processing; ADP Technology Services, Inc.; ADP.com; Automatic Data Processing, Inc.; ADP, Inc.
251dc200b096aaa160b744b7b907b9cc	BD; Becton Dickinson; BD (Becton, Dickinson and Company); Becton Dickinson & Company; Becton Dickinson and Company; Becton, Dickinson & Co.; Becton, Dickinson & Co. (BD); Becton, Dickinson & Company; Becton, Dickinson and Company; Becton, Dickinson; Beckon Dickinson; Beckton Dickinson
9cc691a26109a7474efbe7c0a6f8f066	Coca-Cola; Coca-Cola Bottling; Reyes Coca Cola Bottling; Reyes Coca-Cola Bottling; Reyes Coca-Cola Bottling Group; Coca-Cola Bottling Co. Consolidated; Coca-Cola Bottling Co. Consolidated (CCBCC); Coca-Cola Bottling Company; Coca-Cola Bottling Company Consolidated; Coke consolidated

nyName2Vec. Since company names usually use minimal text, a few words, or a few dozens of characters on average, we first split the company name into characters. We used this representation for a company name instead of a word granularity level. Then used n -gram tokenizer, where n between 1 and 3, and embedded its result using a hash function. Next, we build a sequence embedding sub-model based on a long short-term memory (LSTM) encoder [46] with Rectified Linear Unit (ReLU) activation function [47]. To increase the context available to the algorithm, we used a bidirectional LSTM (Bi-LSTM), a sequence processing model that consists of two LSTMs: one takes the input in a forward direction, and the other in a backward direction. We set character embedding dimensions to 400 and sequence encoder dimensions to 400 as well. We split the company synonyms into two mutually exclusive sets, train and test sets, with a ratio of 9:1, where the train set

consists of about 90% of the synonyms, and the test set 10%. To avoid biasing of the experimental study results (see Chapter 5), we filtered out about 500 names from both the train and the test sets names, which also exist in the dataset we used for measuring the method performance - fortune 1,000 companies dataset (see Section 4.1).

Lastly, we calculated the embedding representation for all company names and synonyms, so it will be possible to calculate using distance metrics (see Section 2.2) the most similar names for each company name.

CHAPTER 4

EXPERIMENTAL STUDY

There are tens of millions of open job ads in the United States published in many job boards and job search engines, like Indeed, ZipRecruiter, Glassdoor, Bing Jobs, and Google Jobs. In this study, we used a job-ads corpus from one of the biggest job boards in the United States, with more than ten million job ads. Although the dataset was acquired from a single source, there was considerable inconsistency in some jobs' properties, such as city names. For example, Los Angeles in California, USA, can be found in many variations, mainly due to letter capitalization, punctuation, and abbreviation, such as "Los-Angeles, CA"; "la, ca"; "los angeles california", and more. Company names are also being regulated and need to comply with the state naming rules in the states where they will be doing business. Otherwise, a state might not accept the documents filed to form or qualify the company. Naming conventions include suffixes like "LLC", "Inc.", "Ltd.", etc.

4.1 Dataset

According to the U.S. Bureau of Labor Statistics (BLS), there are several million job openings in the United States, which increased to a new level of almost ten million job openings after the COVID-19 pandemic hit the markets recently in 2020 [48]. However, looking at job boards and job search engines, like Indeed,¹ ZipRecruiter,² Glassdoor,³ Bing Jobs,⁴ and Google Jobs,⁵ show there are tens of millions of open job ads in the United States published, an order of magnitude larger than what the BLS states. We collected from one of the largest job boards in the United States a large job ads dataset that contains active job ads

¹www.indeed.com

²www.ziprecruiter.com

³www.glassdoor.com

⁴www.bing.com

⁵www.google.com

that existed during June 2020. This dataset includes about ten million job ads. We captured for each job ad the following information: Title, Description, Hiring Company Name, and Location (Country, State, City, and Zip Code).

To evaluate the performance of the described solution, we manually created a ground-truth dataset based on the Fortune 1,000 companies in the United States, where each company has a canonical company name and a list of synonyms. The mean number of synonyms for a canonical company name is 3.8, with a standard deviation of 2.1. This list consists of the most prominent American companies ranked by revenues, compiled and published by Fortune magazine [49]. It only includes companies that are incorporated or authorized to do business in the United States and for which revenues are publicly available regardless of whether they are public companies listed on a stock market.

4.2 Evaluation Process and Performance Metrics

The performance of the recommendations generated by our method was measured by the average of $Success@k$ (Success at rank k) rate, which stands for the mean probability that a relevant company name occurs within the top k of the ranking [50]. Similar to how the relevance of the search engines is measured, in most cases, people are only interested in the first page of the results and do not bother to move on to subsequent pages [51]. In extreme cases, only the top result matters. For example, only a single value can be used when performing a join operation between two tables in a database.

We compared the solution’s performance with edit-distance matching, fuzzy distance matching, and random matching, common best algorithms, as described in Chapter 2. We treated the ground-truth dataset (see Section 4.1) as a mapped list of synonyms to tags, where a tag is the most common company name (canonical name).

We used the ground-truth dataset with the created model described in Chapter 3 and generated a vector representation for both canonical names and synonyms. For each synonym’s vector, we calculated its cosine distance with all the canonical name’s vectors. As a

result, we generated a $N \times M$ matrix R , where N is the total number of canonical names in the ground-truth dataset, and M is the size of the complete set of synonyms. R_{ij} represents the cosine distance between the vector representation of a canonical name i and the vector representation of a synonym j . For each given synonym, we sorted the canonical names based on their cosine distance R_{ij} . Lastly, the closest k canonical names were presented as the recommended canonical names for each given synonym. We used the k closest vectors provided in the matrix R , and calculated the metric of $Success@k$, for $k = 1, 2, 3$. We defined success as equal to 1 in case the canonical company name returned in the closest k company names, or to be equal to 0 otherwise. Namely, we define $Success@k$ as follows:

$$Success@k(synonym) := |\{Tagged\ Canonical\ Name\} \cap \{k\ Closest\ Canonical\ Names\}|$$

Where tagged canonical company name is mapped to the company synonym in the ground-truth dataset.

$$AvgSuccess@k := \frac{\sum_{s \in Synonyms} success@k(s)}{|Synonyms|}$$

CHAPTER 5

RESULTS

To analyze the performance of the implemented method (see Chapter 3), we used the job ads dataset (see Section 4.1) to train the embedding `CompanyName2Vec` model (see Section 3.2), and evaluated the method’s performance using the ground-truth dataset (see Section 4.1). Since the ground-truth dataset is biased to company names which operate in the United States, we first checked the geographical distribution of job corpus dataset. We found that the majority of the job ads, about 88.5%, are posted within the United States.

The job ads fingerprinting result (see Section 3.1) shows that there are 4.25 million of unique job ad fingerprints out of 10 million job ads. From the distribution of job ad fingerprints (see Figure 5.1), only 1.2 million of the unique job ad fingerprints, which were posted with at least two different company names, could be used for training the embedding `CompanyName2Vec` model. We found that applying the agencies staffing companies filters (see Section 3.1) removed 9.5% of the job ads, 8.25% of the fingerprints, and just 2.5% of the company names.

We generated the embedding `CompanyName2Vec` model based on the job ads dataset (see Section 4.1). We used the generated model for resolving CEM and matched the Fortune 1,000 canonical company names with their synonyms. Figure 5.2 presents the distribution of the associated names of the following companies based on the `CompanyName2Vec` vector representation after being reduced into two dimensions using t-SNE [52]: ABM Industries, ACCO Brands, HCA Healthcare, Home Depot, J.B. Hunt Transport Services, Inc., JPMorgan Chase & Co., Lowe’s Inc., and PepsiCo. As can be seen, there are eight clusters, one per every company includes the company synonyms grouped together, which indicates that `CompanyName2Vec` successfully suggested relevant synonyms for those companies.

Finally, we calculated $AvgSuccess@k$ (see Figure 4.2) for the `CompanyName2Vec`

solution and compared it with the following several solutions: random matching, edit-distance matching [20], and fuzzy matching [11]. The `CompanyName2Vec` performs better than all the above solutions, with any given tested k (see Table 5.1). We evaluated this list of algorithms using a 10-fold cross-validation approach. The difference found statistically significant using t-tests with p-value less than 0.001.

Table 5.1: Algorithms’ Performance Comparison

Algorithm	AvgSuccess@1	AvgSuccess@2	AvgSuccess@3
CompanyName2Vec	0.887	0.906	0.910
Fuzzy Distance Matching	0.852	0.857	0.865
Edit Distance Matching	0.565	0.589	0.614
Random Matching	0.001	0.002	0.004

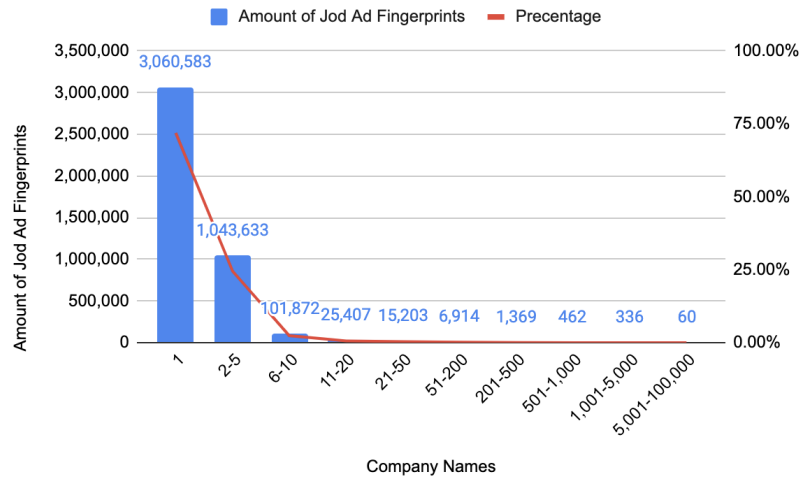


Figure 5.1: Fingerprints Distribution by Company Names

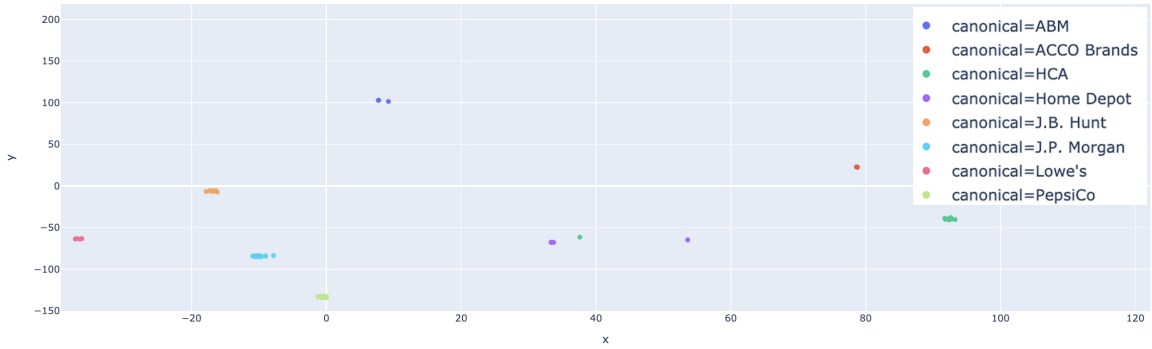
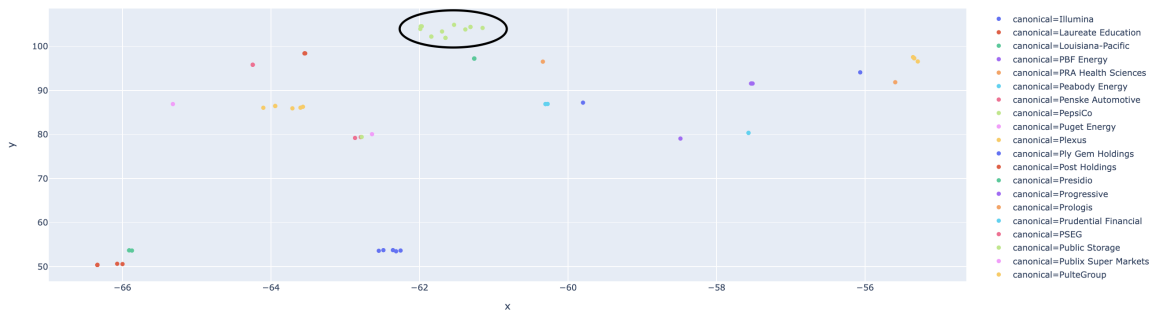
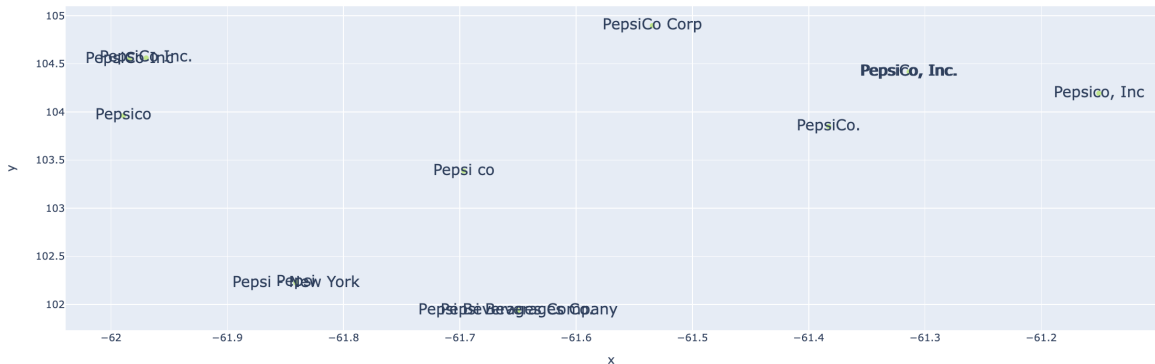


Figure 5.2: Synonym distribution of the following companies as were calculated by the CompanyName2Vec algorithm and reduced into two dimensions with t-SNE: ABM Industries, ACCO Brands, HCA Healthcare, Home Depot, J.B. Hunt Transport Services, Inc., JPMorgan Chase & Co., Lowe's Inc., and PepsiCo

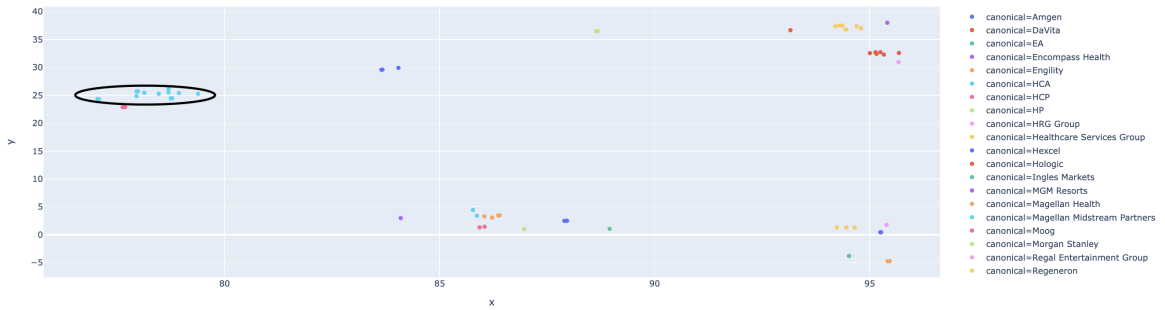


(a) Zoom out view on PepsiCo synonyms and other company synonyms near by

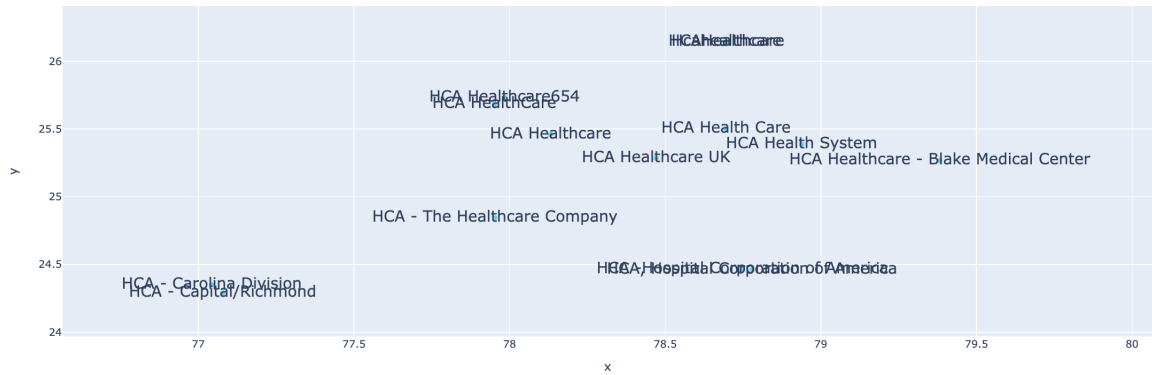


(b) Examples for PepsiCo synonyms

Figure 5.3: A plot of PepsiCo synonyms' vectors after were reduced into two dimensions

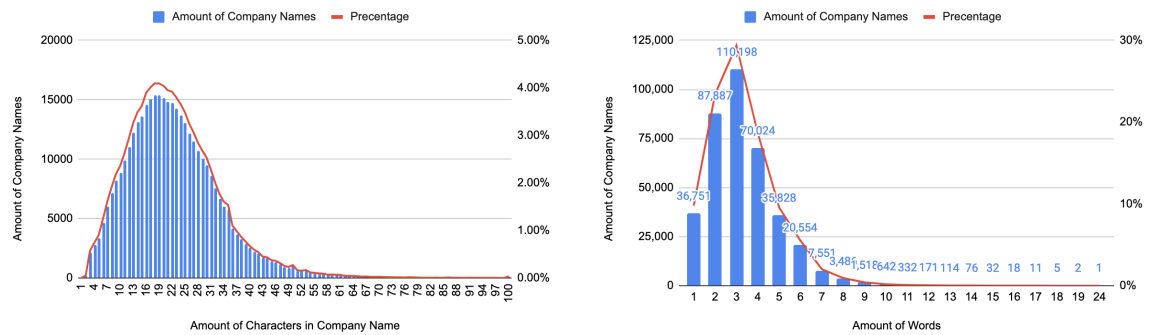


(a) Zoom out view on HCA Healthcare synonyms and other company synonyms near by



(b) Examples for HCA Healthcare synonyms

Figure 5.4: A plot of HCA Healthcare synonyms' vectors after were reduced into two dimensions



(a) Distribution of Company Names by Name's Length

(b) Distribution of company names by amount of words

Figure 5.5: Distribution of company names exist in job ad corpus

CHAPTER 6

DISCUSSION

Upon analyzing the results presented in the previous chapter, we can conclude the following: First, the novel approach we call `CompanyName2Vec`, which defines a vector representation of company names based on job ads, had promising results compared to other algorithms (see Table 5.1) and is helpful for the task of CEM (see Section 4.2). `CompanyName2Vec` algorithm captures the company names semantics and manages common prefixes, suffixes, and punctuation for company synonyms, like PepsiCo's suffixes - Inc, Corp., Co., Company, Etc. (see Figure 5.3). `CompanyName2Vec` manages less popular suffixes as well, which may include a location or a point of interest. For example, HCA Capital Richmond is a synonym of HCA Healthcare. HCA Capital Richmond is a synonym due to the presence of HCA capital offices, which are located in Richmond, VA (see Figure 5.4). Another example is *Pepsi - New York*, which is a synonym of PepsiCo's office in New York City (see Figure 5.3). A company's business line can be attached to a company name as a synonym, such as Home Depot Tools Rental (see Figure A.2), or Lowe's Home Improvement (see Figure A.1). Home Depot Tools Rental is a synonym for Home Depot, and Lowe's Home Improvement is the company slogan.

Second, we observed that "best partial" heuristic (see Section 2.2.1) works well in the case of matching company names because of its short length. The median length of a company name is 21 characters (see Figure 5.5b), and it consists in the median case of just three words (see Figure 5.5b).

Third, based on the method's performance comparison (see Table 5.1), we found the `CompanyName2Vec` outperforms the other algorithms tested, while the second-best algorithm measured is the Fuzzy distance matching.

fourth, as for company abbreviations, due to lack of semantics in company name abbre-

viations, the `CompanyName2Vec` solution performs just as well as Fuzzy distance matching.

Lastly, we found the `CompanyName2Vec` method works better than the Fuzzy method mainly because the `CompanyName2Vec` uses the language semantic. For example, trying to match “Pepsi - New York” with “PepsiCo Inc.” and “New York Life” using Fuzzy distance matching returns “New York Life” because New York is a significant part of both names. The `CompanyName2Vec` method returns “PepsiCo Inc.” based on an LSTM model, which extracts the company naming semantics from the training set.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this study, we introduced `CompanyName2Vec`, a novel, generic open-source algorithm which uses job ads and deep learning to address some of the challenges associated with company name synonyms. We provided a comprehensive description of our algorithm's steps, starting with a collection of job ads and finding company name synonyms using the job ads fingerprinting technique and generated company name synonyms datasets. We used text matching heuristics to reduce wrong synonyms and used the dataset to generate an Bi-LSTM model. This model generates a vector representation given any company name. We collected a set of synonyms for the Fortune 1,000 companies in the United States and used them to test and compare the `CompanyName2Vec` algorithm with several different matching algorithms. We matched all the Fortune 1,000 companies' synonyms with their canonical names and analyzed the results.

We make the following observations and conclusions:

1. The `CompanyName2Vec` algorithm was beneficial for confronting the problem of recommending synonyms for a given company name with higher *Success@k* than other evaluated methods.
2. The `CompanyName2Vec` algorithm helps solve the CEM problem using job-ads data and requires no company's information for matching besides its name.

A possible future research direction is to examine additional configurations that may improve similar name suggestions:

- Using more comprehensive job ads' fingerprinting technique using additional information than just the job description, such as work location, contact information, company logo, Etc.

- Exploring other models than Bi-LSTM for name embedding, like bag of words, convolutional neural network (CNN), and bidirectional encoder representations from transformer (BERT).

Another avenue to pursue is learning company relations from the job ads corpus, for example, a parent company, a subsidiary, a holding company, Etc.

REFERENCES

- [1] Z. Sun, K. Strang, and S. Firmin, “Business analytics-based enterprise information systems,” *Journal of Computer Information Systems*, vol. 57, no. 2, pp. 169–178, 2017.
- [2] J. Liebowitz, “Business analytics and decision-making: The years ahead,” *The World Financial Review*, vol. 28, 2014.
- [3] Z. Sun, L. Sun, and K. Strang, “Big data analytics services for enhancing business intelligence,” *Journal of Computer Information Systems*, vol. 58, no. 2, pp. 162–169, 2018.
- [4] T. Gschwind, C. Miksovic, J. Minder, K. Mirylenka, and P. Scotton, “Fast record linkage for company entities,” in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 623–630.
- [5] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate record detection: A survey,” *IEEE Transactions on knowledge and data engineering*, vol. 19, no. 1, pp. 1–16, 2006.
- [6] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
- [7] D. Vesset, B. McDonough, D. Schubmehl, and M. Wardley, “Worldwide business analytics software 2013–2017 forecast and 2012 vendor shares (doc# 241689),” *Retrieved*, vol. 6, no. 28, p. 2014, 2013.
- [8] M. Paganelli, P. Sottovia, F. Guerra, and Y. Velegrakis, “Tuner: Fine tuning of rule-based entity matchers,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2945–2948.
- [9] L. Getoor and A. Machanavajjhala, “Entity resolution: Theory, practice & open challenges,” *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2018–2019, 2012.
- [10] I. Veldman, “Matching profiles from social network sites,” 2009.
- [11] S. Chaudhuri, V. Ganti, and R. Motwani, “Robust identification of fuzzy duplicates,” in *21st International Conference on Data Engineering (ICDE’05)*, IEEE, 2005, pp. 865–876.
- [12] O. Bousquet, U. von Luxburg, and G. Rätsch, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*. Springer, 2011, vol. 3176.

- [13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “Dbpedia: A nucleus for a web of open data,” in *The semantic web*, Springer, 2007, pp. 722–735.
- [14] K. K. Jones, S. N. Zenk, E. Tarlov, L. M. Powell, S. A. Matthews, and I. Horoi, “A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments,” *BMC research notes*, vol. 10, no. 1, p. 35, 2017.
- [15] L. Getoor and C. P. Diehl, “Link mining: A survey,” *Acm Sigkdd Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.
- [16] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, “Swoosh: A generic approach to entity resolution,” *The VLDB Journal*, vol. 18, no. 1, pp. 255–276, 2009.
- [17] D. G. Brizan and A. U. Tansel, “A. survey of entity resolution and record linkage methodologies,” *Communications of the IIMA*, vol. 6, no. 3, p. 5, 2006.
- [18] I. P. Fellegi and A. B. Sunter, “A theory for record linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.
- [19] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
- [20] G. Navarro, “A guided tour to approximate string matching,” *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.
- [21] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.
- [22] W. E. Yancey, “Evaluating string comparator performance for record linkage,” *Statistics*, vol. 5, p. 38, 2005.
- [23] S. Zhang, Y. Hu, and G. Bian, “Research on string similarity algorithm based on levenshtein distance,” in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, IEEE, 2017, pp. 2247–2251.
- [24] SeatGeek, *Fuzzy-wuzzy python package*, (accessed July 16, 2021).
- [25] W. W. Cohen, “Integration of heterogeneous databases without common domains using queries based on textual similarity,” in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 201–212.

- [26] D. Holmes and M. C. McCabe, “Improving precision and recall for soundex retrieval,” in *Proceedings. International Conference on Information Technology: Coding and Computing*, IEEE, 2002, pp. 22–26.
- [27] A. J. Lait and B. Randell, “An assessment of name matching algorithms,” *Technical Report Series-University of Newcastle Upon Tyne Computing Science*, 1996.
- [28] P. Christen, “A comparison of personal name matching: Techniques and practical issues,” in *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW’06)*, IEEE, 2006, pp. 290–294.
- [29] N. Elbegbayan *et al.*, “Winnowing, a document fingerprinting algorithm,” *TDDC03 Projects, Spring*, 2005.
- [30] S. Schleimer, D. S. Wilkerson, and A. Aiken, “Winnowing: Local algorithms for document fingerprinting,” in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 76–85.
- [31] J. Parapar and Á. Barreiro, “Winnowing-based text clustering,” in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 1353–1354.
- [32] A. T. Wibowo, K. W. Sudarmadi, and A. M. Barmawi, “Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on bahasa indonesia documents,” in *2013 International Conference of Information and Communication Technology (ICoICT)*, IEEE, 2013, pp. 128–133.
- [33] A. Aiken, “Moss: A system for detecting software similarity,” vol. 29, p. 2017, 1994, (accessed July 12, 2021).
- [34] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” in *Advances in Neural Information Processing Systems*, 2001, pp. 932–938.
- [35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [37] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

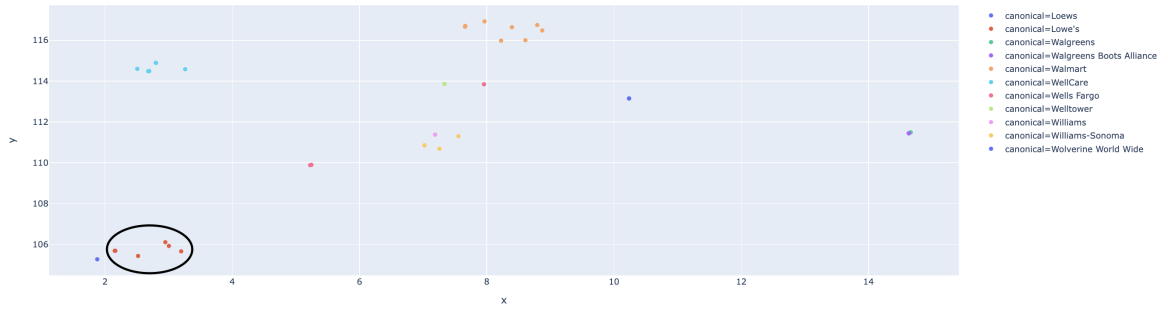
- [38] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “Devise: A deep visual-semantic embedding model,” 2013.
- [39] P. Li, Y. Liu, M. Sun, T. Izuba, and D. Zhang, “A neural reordering model for phrase-based translation,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1897–1907.
- [40] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, “Bilingually-constrained phrase embeddings for machine translation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 111–121.
- [41] J. Zhao, M. Lan, Z.-Y. Niu, and Y. Lu, “Integrating word embeddings and traditional nlp features to measure textual entailment and semantic relatedness of sentence pairs,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2015, pp. 1–7.
- [42] B. Salehi, P. Cook, and T. Baldwin, “A word embedding approach to predicting the compositionality of multiword expressions,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 977–983.
- [43] E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin, “Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning,” *arXiv preprint arXiv:1509.01692*, 2015.
- [44] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.
- [45] R. Rivest and S. Dusse, *The md5 message-digest algorithm*, 1992.
- [46] K. Greff, R. K. Srivastava, J. Koutnk, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [47] A. F. Agarap, “Deep learning using rectified linear units (relu),” *arXiv preprint arXiv:1803.08375*, 2018.
- [48] U. B. of Labor Statistics, *Job openings and labor turnover summary*, (accessed June 21, 2021).
- [49] *Fortune magazine*, (accessed November 14, 2021).

- [50] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber, “Offline and online evaluation of news recommender systems at swissinfo. ch,” in *Proceedings of the 8th ACM Conference on Recommender systems*, 2014, pp. 169–176.
- [51] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.
- [52] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

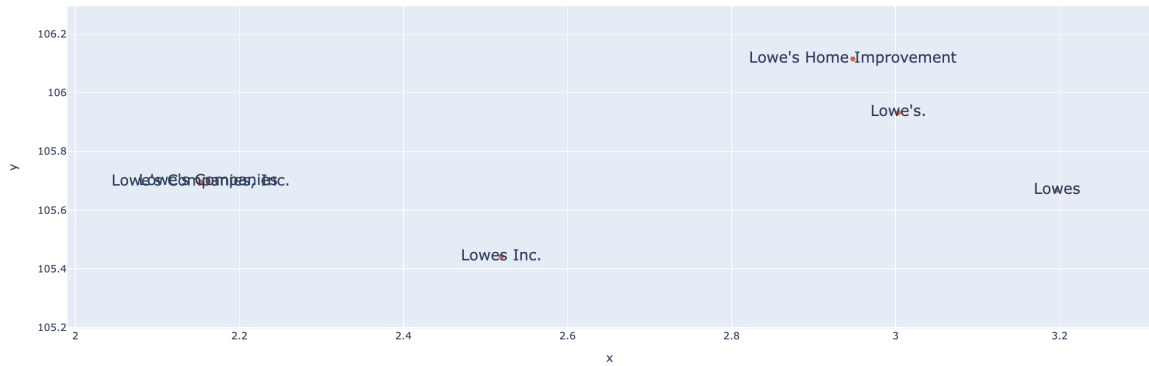
Appendices

APPENDIX A

ADDITIONAL RESULTS

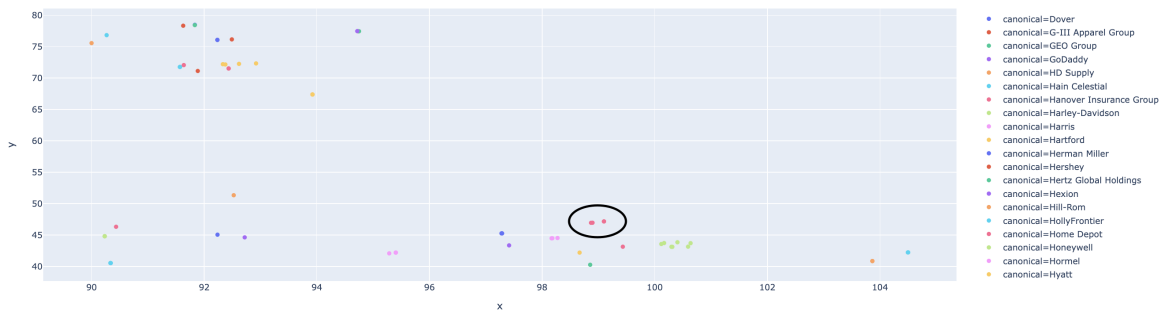


(a) Zoom out view on Lowe's Inc. synonyms and other company synonyms near by

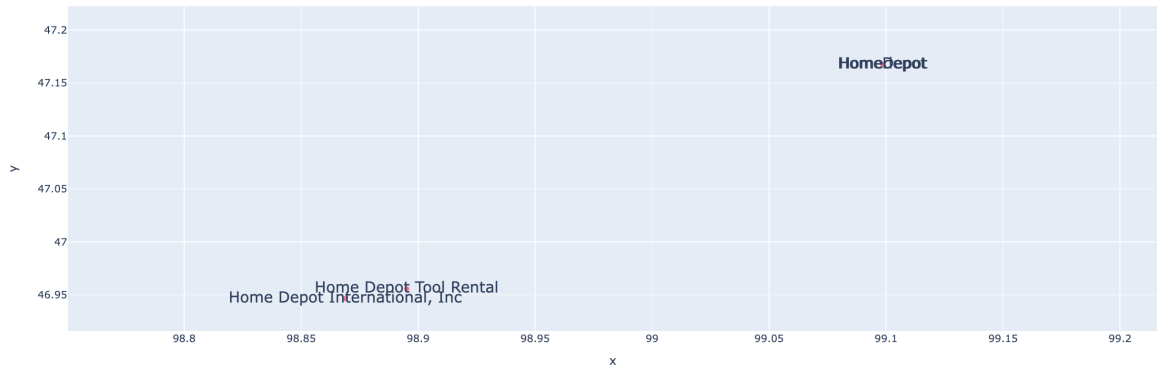


(b) Examples for Lowe's Inc. synonyms

Figure A.1: A plot of Lowe's Inc. synonyms' vectors after were reduced into two dimensions



(a) Zoom out view on Home Depot synonyms and other company synonyms near by



(b) Examples for Home Depot synonyms

Figure A.2: A plot of Home Depot synonyms' vectors after were reduced into two dimensions

APPENDIX B
REFERENCES TO PROJECT CODE AND DATASETS

Code and datasets can be found at the following link:

<https://tinyurl.com/he4tzk>



Reichman University

Efi Arazi School of Computer Science

M.Sc. program - Research Track

CompanyName2Vec: Company Entity Matching Based on Job Ads

by

Ran Ziv

M.Sc. dissertation, submitted in partial fulfillment of the requirements for the M.Sc.
degree, research track, School of Computer Science, Reichman University

September 2021

This work was carried out under the supervision of Dr. Michael Fire from the department of Software and Information Systems Engineering, Faculty of Engineering, Ben Gurion University, and with guidance from Dr. Ilan Gronau from the Efi Arazi School of Computer Science, Reichman University

Abstract

Entity Matching is an essential part of all real-world systems that take in structured and unstructured data coming from different sources. Typically no common key is available for connecting records. Massive data cleaning and integration processes require completion before any data analytics or further processing can be performed. Although record linkage is frequently regarded as a somewhat tedious but necessary step, it reveals valuable insights, supports data visualization, and guides further analytic approaches to the data. Here, we focus on organization entity matching. We introduce `CompanyName2Vec`, a novel algorithm to solve company entity matching (CEM) using a neural network model to learn company name semantics from a job ad corpus, without relying on any information on the matched company besides its name. Based on a real-world ground truth dataset, we show that `CompanyName2Vec` outperforms other evaluated methods and solves CEM challenge with average success rate of 89.3%

TABLE OF CONTENTS

List of Tables	vi
List of Figures	vii
Chapter 1: Introduction	1
Chapter 2: Related Work	5
2.1 Entity Matching	5
2.1.1 Supervised Approaches	5
2.1.2 Unsupervised Approaches	6
2.2 Distance Metrics	6
2.2.1 Character-Based Similarity Metrics	7
2.2.2 Token-Based Similarity Metrics	8
2.2.3 Phonetic Similarity Metrics	8
2.3 Document Fingerprinting	9
2.4 Text Embedding	9
Chapter 3: Method	11
3.1 Hiring Company Names and Synonyms	12
3.2 Job Ad Fingerprinting	12

3.3	Company Synonyms	13
3.4	Company Name Embedding	13
	Chapter 4: Experimental Study	16
4.1	Dataset	16
4.2	Evaluation Process and Performance Metrics	17
	Chapter 5: Results	19
	Chapter 6: Discussion	26
	Chapter 7: Conclusions and Future Work	28
	References	30
	Appendices	34
	Appendix A: References to project Code and Datasets	35

LIST OF TABLES

2.1	Company Names Synonyms Ratio Similarity vs. Partial Ratio Similarity . .	8
3.1	For example, we provide four fingerprint values and the lists of hiring company names posted with the same job ad's fingerprint	14
5.1	Algorithms' Performance Comparison	20

LIST OF FIGURES

3.1	Flowchart of the implemented methodology.	11
5.1	Fingerprints Distribution by Company Names	20
5.2	Synonym distribution of the following companies as were calculated by the CompanyName2Vec algorithm and reduced into two dimensions with t-SNE: ABM Industries, ACCO Brands, HCA Healthcare, Home Depot, J.B. Hunt Transport Services, Inc., JPMorgan Chase & Co., Lowe’s Inc., and PepsiCo	21
5.3	A plot of PepsiCo synonyms’ vectors after were reduced into two dimensions	21
5.4	A plot of HCA Healthcare synonyms’ vectors after were reduced into two dimensions	22
5.5	A plot of Lowe’s Inc. synonyms’ vectors after were reduced into two dimensions	23
5.6	A plot of Home Depot synonyms’ vectors after were reduced into two dimensions	24
5.7	Distribution of company names exist in job ad corpus	25

CHAPTER 1

INTRODUCTION

Enterprise Business intelligence systems have emerged as a disruptive technology, and innovative solution to the global economy [1]. Business intelligence became an emerging and fast-growing field in the past years [2]. Big data and its emerging technologies, including business intelligence and big data analytics systems, have become a mainstream market adopted broadly across industries, organizations, and geographic regions to facilitate data-driven decision making and significantly affecting the way that decision-makers, such as CEOs, operate and run their business [2, 3]. One of the fundamental functionalities a business intelligence system must have is the ability to integrate many data sources [2, 4]. The integration of multiple sources usually requires linking between the significant entities that exist across data sources and systems. Usually, those entities function as a “primary key” in each system [4]. The technique used for performing such linkage is commonly referred to as “Record Linkage,” “Data Deduplication,” “Object Matching,” or “Entity Matching” [4, 5].

Entity Matching (EM) is a fundamental task in data integration scenarios. EM is the task to identify semantically equivalent entities referring to the same real-world object (e.g., persons, products, companies) within one data source or between different sources. EM is also a core technique for data cleaning [5, 6] and data integration [7, 8]. Accurate and fast entity matching has huge practical implications in a wide variety of commercial, scientific and security domains [9]. EM solutions can be divided into three separate groups: supervised, unsupervised, and semi-supervised approaches [10, 11]. Supervised approaches require labeled training sets or predefined thresholds on which to base their decisions. Unfortunately, in most real-world cases, the variety of patterns that can be observed are not feasible to be captured in a training set, and therefore, these solutions are quite limited.

Unsupervised approaches help find previously unknown patterns in a dataset without pre-existing labels and are based on clustering algorithms that group together items with high similarity [12]. Semi-supervised approaches fall between the unsupervised approach (with no labeled training data) and a supervised approach (with only labeled training data). Semi-supervised approaches are based on a small set of labeled data and a large set of unlabeled data.

Enterprise business intelligence applications require the integration of many data sources. In such applications, one of the most crucial entity attributes to be linked is the company name [4]. A company usually exists in many of the enterprise systems. For example, a company can represent a potential customer or a sales lead in the sales system, an existing customer in the customer relationship management system, a supplier in the logistics system, etc. Although a company name can be noisy, in most cases, this is one of the most potent properties for company entity matching (CEM) due to the lack of unique shared identifiers across datasets [4]. To improve CEM results, many of the existing methods use more company properties other than the company name, such as location, primary phone number, industry, website URL, and more. Such company information is usually costly to acquire on a large scale. The largest public source for such company information is DBpedia [13]. DBpedia contains about 65,000 company entities worldwide, derived from the English version of Wikipedia [4]. There are several companies worldwide that hold a much larger dataset of company information, such as Dun and Bradstreet and Infogroup [14]. These companies do not provide the complete dataset but provide a limited paid service for company information enrichment, mainly for sales and marketing purposes.

In this study, we developed `CompanyName2Vec` (see Chapter 3), a novel algorithm to solve CEM using a neural network model to learn company name semantics from a job ad corpus. Once trained, such a model can suggest synonymous company names. As the name implies, `CompanyName2Vec` represents each different company name as a vector. The vectors are chosen carefully such that a simple mathematical function, like

cosine similarity between vectors, indicates the level of semantic similarity between the company names represented by those vectors. We used `CompanyName2Vec` for CEM. For this purpose, we created a real-world dataset that consists of the largest companies in the US and their synonyms. We used this dataset to evaluate the `CompanyName2Vec` algorithm and compared it with other known entity matching methods.

To develop the `CompanyName2Vec` algorithm, we needed to have a relatively large corpus of company names. Job ads consist company names, and are available publicly on a large scale but are very noisy (see Section 3.3) due to how job ad distribution works. Companies distribute their jobs to multiple job boards and sometimes to recruitment and staffing agencies to reach as many job seekers as possible. In many cases, the process for posting a job ad on job boards requires the employer to manually type the job information, including company name, location, title, description, and more, which causes some inconsistency and duplication of job ads. For example, Tesla’s job ad can be posted with several employer names like `tesla`; `Tesla, Inc.`; `Tesla Motors`; `Tesla Motors, Inc.`; and `Ursus`, which is a staffing and recruitment agency. `CompanyName2Vec` leverages the availability of job ads publicly (see Section 4.1) and their inconsistency to learn company name semantics (see Section 3.4).

Our method includes the following steps: First, we collected a large dataset of job ads (see Section 4.1). Second, we used a fingerprinting technique and created a list of company names that posted the same job (see Sections 3.2 and 3.1). A data cleansing step was also required to filter out non-valid job ads and company names to create a ground truth dataset. Next, we pre-processed this cleaned dataset and created train and test sets (see Section 3.3). Afterward, we evaluated a neural network architecture to create a vector representation for a company name (see Section 3.4). We developed a method to learn what parts in the company name are more important than others and created a vector representation that reflects it accordingly. For example, common suffixes like `Inc.` and `LLC`, usually define the company legal entity type, and employers sometimes do not include them when posting a job

ad because of their low importance. Subsequently, we used cosine distance for measuring the similarity between different vectors (see Section 3.4). Lastly, we created a real-world dataset to evaluate the complete CEM solution (see Chapter 4).

An example for an application of this method is optimizing a company's sales operations, where there is a need to filter irrelevant sale leads related to existing customers. In order to do so, it is required to join two tables, table A exists in the marketing system and includes sales leads, and the other table B lives in the billing system and includes the company's existing customers. Due to the lack of a cross-systems unique identifier, the only way to join those tables is by matching the company names. The described algorithm resolves such a join by using the embedding model to create a vector representation for every company name in tables A and B and finding each company name in table A the similar company name in table B by utilizing the vectors distance (see Section 3.4).

Our main contributions are twofold:

- `CompanyName2Vec`, a novel algorithm to create a mathematical representation for a company name that holds the naming semantics, which enables it to measure similarity between any given company names without relying on any information on the matched company besides its name - available in Chapter 3.
- An end-to-end, highly scalable, enterprise-grade system that uses `CompanyName2Vec` algorithm to suggest company name synonyms to solve company entity matching - available in Chapter 4.

The remainder of the paper is organized as follows: In Chapter 2, we present an overview of related studies. In Chapter 3, we present the methods used to create the `CompanyName2Vec` embedding model. In Chapter 4, we describe the experimental study we conducted to construct the model and evaluate it. Chapter 5, we present the obtained results. In Chapter 6, we discuss the obtained results. Lastly, in Chapter 7, we present our conclusions from this study and offer future research directions.

CHAPTER 2

RELATED WORK

Entity Matching (EM) is an important task, which many researchers in the past have addressed [5, 15, 16, 17]. In this study, we addressed a particular case of EM, namely, the matching of company names across multiple datasets and systems. This chapter describes various related research results and discusses their connection to the solution we developed.

2.1 Entity Matching

EM addresses matching entities between different data sources or deduplication of entities within a single source. EM uses algorithms to both detect duplicates in data and resolve them. As described above, EM solutions can be divided into three separate groups: supervised, unsupervised, and semi-supervised approaches. In the following subsections, we describe the different methods for EM.

2.1.1 Supervised Approaches

The supervised approaches are based on training data in the form of record pairs, pre-labeled as match or not match. In 1969, Ivan Fellegi and Alan Sunter [18] denote two classes: A class M , which represents matches, and a class U , which represents non-matches. There are three main approaches: Rule-based, learning-based, and distance-based.

Rule-Based. One way of performing entity matching is by setting a set of rules that identify the conditions that would make two records be considered as matched. For instance, matching a customer entity which includes its name and full address, it could be of the form: $(name, edit, =, 1) \wedge (address, edit, >, 0.7)$, which indicates that two customer records would match if their names are fully matched and their address and city attributes are more than 70% similar.

Learning-Based. Learning-based approaches [5] use training sets that consist of pairs labeled as Match or Non-match. Each pair includes a comparison vector that represents the comparable attributes of the two items in the pair. Assuming the density function for the classes M and U are different, the EM problem can be treated as a Bayesian inference problem.

Distance-Based. Distance-based approaches [5] theoretically do not need labeled data. In these approaches, a distance metric is defined between data items. A decision is made based on whether or not this pair is a match or not based on the distance between two items and threshold. This threshold can be set by making a reliable estimation. A good threshold can improve the results, therefore using a training set for setting this threshold might be desirable.

2.1.2 Unsupervised Approaches

The idea behind unsupervised approaches is that similar comparison vectors correspond to the same class [5]. Unsupervised learning for EM has its roots in the probabilistic model proposed in 1969 by Ivan Fellegi and Alan Sunter [18]. When there is no training data to compute the probability estimates, it is possible to use variations of the Expectation-Maximization algorithm [19] to better identify appropriate clusters in the data.

2.2 Distance Metrics

A common source for mismatches in database entries is the typographical variation of string data. One of the methods to deal with typographical variations is to measure the similarity between two different strings. In this section, we describe string matching techniques that have been applied in EM.

2.2.1 Character-Based Similarity Metrics

The character-based similarity metrics method is designed to efficiently manage typographical errors. A popular character-based method is the Levenshtein distance (also referred as Edit Distance) [20]. This method measures similarity between two strings by counting the minimum number of operations, including insertion, deletion or substitutions, required to transform one string into the other. Damerau-Levenshtein distance [20, 21] is a variation of the Levenshtein distance where a transposition of two characters is also considered to be an elementary edit operation in addition to insertion, deletion, and substitution. Jaro distance [22] is based on the number of common characters and the number of transpositions in two strings. Winkler distance [22] is a variation of the Jaro distance and gives higher scores to strings that share the same prefix.

In terms of company names, Levenshtein Distance seems to work well for a single word or a much longer text, but not for just for a few words [23], as in our use case. For example, “New York Yankees” and “Yankees” are clearly referring to the same company name, but “New York Mets” and “New York Yankees” are clearly referring to different ones. Yet, the score of the “wrong” match is higher than the “right” one. For such cases there is an heuristic called “best partial” [24], which uses partial matching logic. This logic is implemented in Fuzzy-Wuzzy package [24], which based on Levenshtein Distance [20] to calculate the differences between two different strings. Given two strings X and Y , let the shorter string X be of length m . It finds the ratio similarity measure between the shorter string X and every substring of length m in the longer string Y , and returns the maximum of those similarity measures. So in the case of “Yankees” and “New York Yankees” the score will be higher than the score of “New York Mets” and “New York Yankees” (see Table 2.1), since the substring “Yankees” is completely contained in the string “New York Yankees”.

Table 2.1: Company Names Synonyms Ratio Similarity vs. Partial Ratio Similarity

String X	String Y	Ratio Similarity	Partial Ratio Similarity
YANKEES	NEW YORK YANKEES	61%	100%
NEW YORK METS	NEW YORK YANKEES	76%	69%
NEW YORK METS	NEW YORK MEATS	96%	92%

2.2.2 Token-Based Similarity Metrics

The method of Character-based similarity metrics measures similarity depending only on the appearance and sequence of characters, while token-based similarity metrics first tokenize two strings into sets of tokens (words) and only then compute the similarity between the two sets. Token-based similarity metrics are robust in measuring the similarity of full names, for example, where the order of the first name and last name may change from one string to another. A standard method for Token-based similarity metrics is *TF-IDF* (Term Frequency / Inverted Document Frequency) [25], a numerical statistic method that intends to reflect how important a word is to a document in a collection of documents. *Cosine similarity* [25] measures the similarity of strings by transforming words into vectors, where the frequency of a word is a dimension in the vector. Cosine similarity measures the similarity between two strings by measuring the angle between the vectors. Token-based similarity metrics methods do not take misspellings into account. Therefore those methods are usually used together with character-based methods to determine whether two tokens are similar enough [25].

2.2.3 Phonetic Similarity Metrics

Strings may be phonetically similar even if they are not similar at the character or token level. Different from character-based and token-based similarity metrics, phonetic similarity metrics are limited to a string-based representation. An example of this is, the name

Claire. It has two alternatives, Clare and Clair, which are both pronounced the same. Soundex [26, 27], for example, is one of the best known phonetic encoding algorithms for indexing names by sound as pronounced in English. Soundex keeps the first letter and converts the rest of the string into numbers according to a phonetic encoding table. Additional string comparison methods can be found in a thorough survey written by Peter Christen [28].

2.3 Document Fingerprinting

Among digital data, documents are the easiest to copy and remove any signatures or fingerprints embedded, which make the pirating the hardest to detect [29]. Anyone can just retype a document or copy a part of it. Document fingerprinting is concerned with accurately identifying and copying, including small partial copies, within large sets of documents [29]. In our study, we use fingerprinting for finding hiring company synonyms in the job ads dataset by calculating a job ad fingerprint and looking for hiring company names with the same fingerprint. Checksum for example is a small digest of the full document. This method is simple and sufficient for detecting exact copies, but in some cases, as well as for our study, there is a requirement in a method which is more local and detects partial copies like the Winnowing algorithm [30], which selects the q -gram whose hash value is the minimum within a sliding window of q -grams. Winnowing is used for text clustering [31], and detecting plagiarism, like the comparison Agung Toto Wibowo published in 2013 [32] which detects plagiarism fraud on Bahasa Indonesia documents, or Moss by Alex Aiken [33] which is a system for measuring software similarity, used by Stanford university for detecting plagiarism in programming classes.

2.4 Text Embedding

In this study, we used embedding to create a mathematical representation for a company name that holds the naming semantics. When some object x is said to be embedded in

another object y , the embedding is given by some injective and structure-preserving map $f : X \rightarrow Y$. We used text embedding, similarly to what Yoshua Bengio proposed in 2001 by [34], in the form of a feed-forward neural network language model. Modern methods use a simpler and more efficient neural architecture to learn word vectors, like word2vec [35, 36], and GloVe [37], based on objective functions that are explicitly designed to produce high-quality vectors. Neural embedding learned by these methods have been applied in a myriad of NLP applications, including initializing neural network models for objective visual recognition [38], or machine translation [39, 40], as well as directly modeling word-to-word relationships [35, 41, 42, 43], Paragraph vectors, or doc2vec, were proposed in 2014 by Quoc Le and Tomas Mikolov [44] as a simple extension to word2vec to extend the learning of embedding from words to word sequences. Doc2vec¹ is agnostic to the granularity of the word sequence, and it can equally be a word n-gram, sentence, paragraph, or document. One of the benefits of using dense and low-dimensional vectors is computational. The main benefit of this dense representation is generalization power. If we believe some features may provide similar clues, it is worthwhile to provide a representation that can capture these similarities. We used text embedding to learn company names semantics and to measure similarity.

¹The term doc2vec was popularized by Gensim, a widely-used implementation of paragraph vectors: <https://radimrehurek.com/gensim/>

CHAPTER 3

METHOD

The method described below deals with CEM by creating a generic open solution that matches company names. Unlike many other existing solutions that utilize expensive commercial datasets, the presented method's novelty uses job-ads data, and it requires no company's information for matching besides its name.

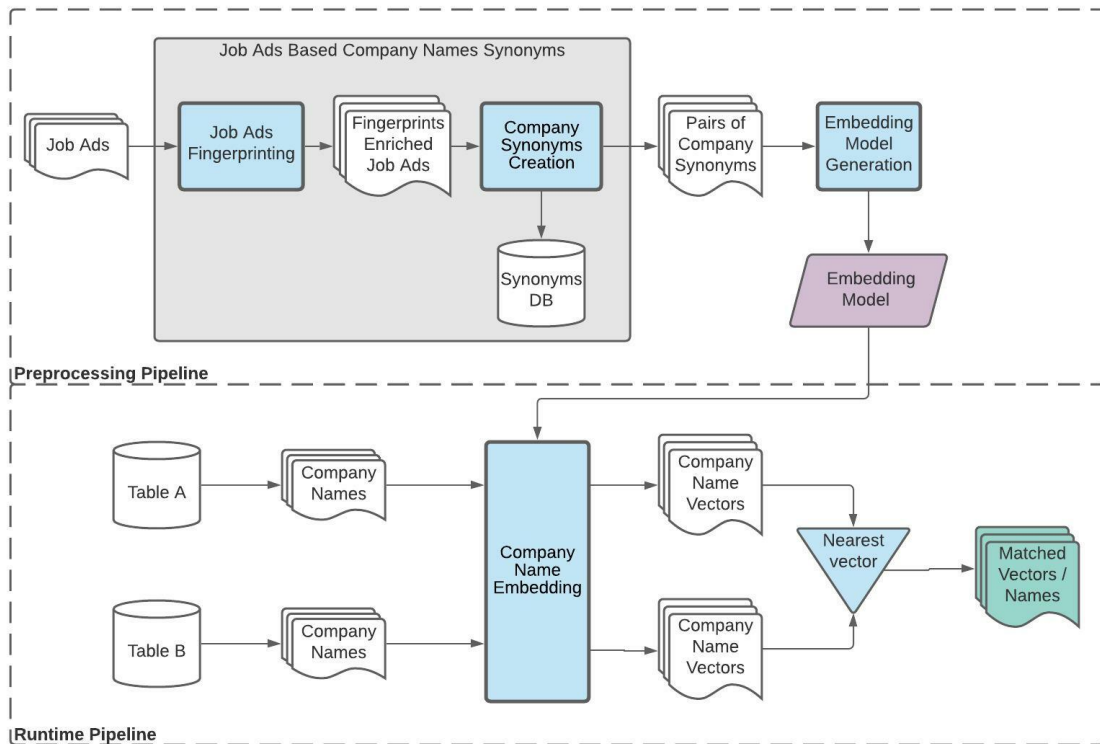


Figure 3.1: Flowchart of the implemented methodology.

Our method consists of two main parts: first, the creation of pairs, in which a pair consists of synonyms for the same company based on an analysis of large-scale job ads data (see Section 3.1); second, the algorithm `CompanyName2Vec` for constructing an embedding model that utilizes these pairs of synonyms. The embedding model takes as input a company name and returns as output meaningful vector representation, where two company

names that represent the same company will have similar vectors (see Section 3.4).

3.1 Hiring Company Names and Synonyms

Given a job ads corpus, we first create and enrich each job ad with a unique job ad identifier (aka job ad fingerprint) using a fingerprinting technique (see Section 3.2). Second, we use the job ad fingerprint to identify company name synonyms. We group the job ad data by job ad fingerprint and create a list of company names posted with the same fingerprint, or in other words, company names that were published within the same job ad. Then, we remove duplicated synonyms and create a list of pairs, where each pair includes company names that were posted within one or more job ads (see Section 3.3).

A particular case we need to handle is staffing and recruitment agencies, which often publish their customers' job ads anonymously and use the staffing or recruitment agency name instead. Ideally, each pair includes company names that refer to the same company. Therefore, to reduce such noise, we filter out staffing and recruitment agencies' names. (see Section 3.3).

3.2 Job Ad Fingerprinting

Given a dataset of job ads, our method's first step is to create a unique identifier (aka fingerprint) for each job ad. This component calculates a job ad fingerprint (see Section 2.3) based on the job description. We used a local approach to calculate fingerprints [30]. First we calculated a Winoing list of fingerprints. We used the following parameters: $kgram_len = 4$, $window_len = 5$, $base = 10$ and $modulo = 1000$. Then, we utilized the MD5 hashing algorithm [45] to convert the list of fingerprints into a single short 128-bit job identifier to improve computation and debugging efficiency.

3.3 Company Synonyms

The purpose of collecting company synonyms from job ads is to create a dataset for the method’s embedding model (see Section 3.4). An outcome of it is the creation of a large database of company synonyms. We utilize the fingerprints to group company names with the same job ad’s fingerprint to construct training and testing sets. After grouping the company names by fingerprint, it transforms the group of names into a set of pairs (i_{name}, j_{name}) where i_{name} ’s and j_{name} ’s Job Ads have at least one Job Ad’s fingerprint in common. Namely, let A be a set of job ads, and let F be a set of fingerprints. The function $Fingerprint : A \mapsto F$ maps a job adds to a fingerprint. Then, for each fingerprint value a , we define the following set:

$$\forall i, j \in A, (i.name, j.name) \in F_a \iff Fingerprint(i) = Fingerprint(j)$$

It is important to note that a company may have several fingerprints, at least as many as the number of open positions it tries to fill. In addition, we applied a filter to remove staffing and recruitment agencies using naive text matching. From our analysis, we found that much of the noise can be reduced by filtering company names which include the following strings: “staff,” “recruit,” “jobs,” or “unknown.”

Lastly, we removed duplicated pairs from this large list of names pairs, which created when two company names have more than just a single job ad fingerprint in common.

3.4 Company Name Embedding

In the algorithm’s second step, we utilize an embedding algorithm (see Section 2.4) to capture the semantics of company names by placing semantically similar company names close together in the embedding space. We call this embedding solution `CompanyName2Vec`. Since company names usually use minimal text, a few words, or a few dozens of characters

Table 3.1: For example, we provide four fingerprint values and the lists of hiring company names posted with the same job ad’s fingerprint

Fingerprint	Company Names
9abc97a34a 7c5630b0673083fb9c61c1	99 Cents Only Stores; 99 Cent Only; 99 Cents Only; 99 Cents Only Stores LLC A.O. Smith; AO Smith; AO Smith Corporation; A. O. Smith; A. O. Smith Corporation; A. O. smith; Smith (A.O.) Corporation; Next Generation Recruitment and Staffing Agency
c9245756a0edca343c96a1a3b8762fc0	ADP; ADP Automatic Data Processing; ADP Technology Services, Inc.; ADP.com; Automatic Data Processing, Inc.; ADP, Inc.
251dc200b096aaa160b744b7b907b9cc	BD; Becton Dickinson; BD (Becton, Dickinson and Company); Becton Dickinson & Company; Becton Dickinson and Company; Becton, Dickinson & Co.; Becton, Dickinson & Co. (BD); Becton, Dickinson & Company; Becton, Dickinson and Company; Becton, Dickinson; Beckon Dickinson; Beckton Dickinson
9cc691a26109a7474efbe7c0a6f8f066	Coca-Cola; Coca-Cola Bottling; Reyes Coca Cola Bottling; Reyes Coca-Cola Bottling; Reyes Coca-Cola Bottling Group; Coca-Cola Bottling Co. Consolidated; Coca-Cola Bottling Co. Consolidated (CCBCC); Coca-Cola Bottling Company; Coca-Cola Bottling Company Consolidated; Coke consolidated; Leap Staffing Agency

on average, we first split the company name into characters and used this representation for a company name instead of a word granularity level. Then used n-gram tokenizer, where n between 1 and 3, and embedded its result using a hash function. Next, we build a sequence embedding sub-model based on a long short-term memory (LSTM) encoder [46] with Rectified Linear Unit (ReLU) activation function [47]. In order to increase the context available to the algorithm, we used a bidirectional LSTM (Bi-LSTM), a sequence processing model that consists of two LSTMs: one takes the input in a forward direction, and the other in a backward direction. We set character embedding dimensions to 400 and sequence encoder dimensions to 400 as well. We split the company synonyms into two mutually exclusive

sets, train and test sets, with a ratio of 9:1, where the train set consists of about 90% of the synonyms, and the test set 10%. In order to avoid biasing of the experimental study results (see 5), we filtered out about 500 names from both the train and the test sets names, which also exist in the dataset we used for measuring the method performance - fortune 1,000 companies dataset (see 4.1).

Lastly, we calculated the embedding representation for all company names and synonyms, so it will be possible to calculate using distance metrics (see Section 2.2) the most similar names for each company name.

CHAPTER 4

EXPERIMENTAL STUDY

There are tens of millions of open job ads in the United States published in many job boards and job search engines, like Indeed, ZipRecruiter, Glassdoor, Bing Jobs, and Google Jobs. In this study, we used a job-ads corpus from one of the biggest job boards in the United States, with more than ten million job ads. Although the dataset was acquired from a single source, there was considerable inconsistency in some job’s properties, such as city names. For example, the city Los Angeles in California, USA can be found in many variations, mainly due to letter capitalization, punctuation and abbreviation, such as “Los-Angeles, CA”; “la, ca”; “los angeles california,” and more. Company names are also being regulated and need to comply with the state naming rules in the states where they will be doing business. Otherwise, a state might not accept the documents filed to form or qualify the company. Naming conventions include suffixes like “LLC,” “Inc.,” “Ltd.,” etc.

4.1 Dataset

According to the U.S. Bureau of Labor Statistics (BLS) there are several million job openings in the United States, which increased to a new level of almost ten million job openings after the COVID-19 pandemic hit the markets recently in 2020 [48]. However, looking at job boards and job search engines, like Indeed,¹ ZipRecruiter,² Glassdoor,³ Bing Jobs,⁴ and Google Jobs,⁵ show there are tens of millions of open job ads in the United States published, an order of magnitude larger than what the BLS states. We collected from one of the largest job boards in the United States a large job ads dataset that contains active job ads

¹www.indeed.com

²www.ziprecruiter.com

³www.glassdoor.com

⁴www.bing.com

⁵www.google.com

that were existed during June 2020. This dataset includes about ten million job ads. We captured for each job ad the following information: Title, Description, Hiring Company Name, and Location (Country, State, City, and Zip Code).

To evaluate the performance of the described solution, we manually created a ground-truth dataset for company canonical names and their synonyms based on the Fortune 1,000 companies in the United States. This list consists the largest American companies ranked by revenues, as compiled and published by the American business magazine Fortune. It only includes companies that are incorporated or authorized to do business in the United States and for which revenues are publicly available regardless of whether they are public companies listed on a stock market.

4.2 Evaluation Process and Performance Metrics

The performance of the recommendations generated by our method was measured by the average of *Success@k* (Success at rank k) rate, which stands for the mean probability that a relevant company name occurs within the top k of the ranking. Similar to how the relevance of the search engines is measured, in most cases, people are only interested in the first page of the results and do not bother to move on to subsequent pages [49]. In extreme cases, only the top result matters. For example, only a single value can be used when performing a join operation between two tables in a database.

We compared the solution’s performance with edit-distance matching, fuzzy distance matching, and random matching, common best algorithms, as described in Chapter 2. We treated the ground-truth dataset (see Section 4.1) as a mapped list of synonyms to tags, where a tag is the most common company name (canonical name).

We used the ground-truth dataset with the created model described in Chapter 3 and generated a vector representation for both canonical names and synonyms. For each synonym’s vector, we calculated its cosine distance with all the canonical name’s vectors. As a result, we generated a $N \times M$ matrix R , where N is the total number of canonical names in

the ground-truth dataset, and M is the size of the complete set of synonyms. R_{ij} represents the cosine distance between the vector representation of a canonical name i and the vector representation of a synonym j . For each given synonym, we sorted the canonical names based on their cosine distance R_{ij} . Lastly, the closest k canonical names were presented as the recommended canonical names for each given synonym. We used the k closest vectors provided in the matrix R , and calculated the metric of $Success@k$, for $k = 1, 2, 3$. We defined success as equal to 1 in case the canonical company name returned in the closest k company names, or to be equal to 0 otherwise. Namely, we define $Success@k$ as follows:

$$Success@k(synonym) := |\{Tagged\ Canonical\ Name\} \cap \{k\ Closest\ Canonical\ Names\}|$$

Where tagged canonical company name is mapped to the company synonym in the ground-truth dataset.

$$AvgSuccess@k := \frac{\sum_{s \in Synonyms} success@k(s)}{|Synonyms|}$$

CHAPTER 5

RESULTS

To analyze the performance of the implemented method (see Chapter 3), we used the job ads dataset (see Section 4.1) to train the embedding `CompanyName2Vec` model (see Section 3.4). First, we analyzed the job ads dataset. We found that the majority of the job ads, about 88.5%, are posted within the United States, similarly to the ground-truth dataset (see Section 4.1), which was used in the solution’s evaluation process.

The job ads fingerprinting result (see Section 3.2) shows that there are 4.25 million of unique job ad fingerprints out of 10 million job ads. From the distribution of job ads fingerprints (see Figure 5.1), only 1.2 million of the unique job ad fingerprints, which were posted with at least two different company names, could be used for training the embedding `CompanyName2Vec` model. We found that applying the agencies staffing companies filters (see Section 3.3) removed 9.5% of the job ads, 8.25% of the fingerprints, and just 2.5% of the company names.

We generated the embedding `CompanyName2Vec` model based on the job ads dataset (see Section 4.1). We used the generated model for resolving CEM and matched the Fortune 1,000 canonical company names with their synonyms. Figure 5.2 presents the distribution of the associated names of the following companies based on the `CompanyName2Vec` vector representation after being reduced into two dimensions using t-SNE: ABM Industries, ACCO Brands, HCA Healthcare, Home Depot, J.B. Hunt Transport Services, Inc., JPMorgan Chase & Co., Lowe’s Inc., and PepsiCo. As can be seen, there are eight clusters, one per every company includes the company synonyms grouped together, which indicates that `CompanyName2Vec` successfully suggested relevant synonyms for those companies.

Finally, we calculated *AvgSuccess@k* (see Figure 4.2) for the `CompanyName2Vec` solution and compared it with the following several solutions: random matching, edit-

distance matching [20], and fuzzy matching [11]. The `CompanyName2Vec` performs better than all the above solutions, with any given tested k (see Table 5.1). The difference found statistically significant using t-tests with p-value less than 0.001.

Table 5.1: Algorithms’ Performance Comparison

Algorithm	AvgSuccess@1	AvgSuccess@2	AvgSuccess@3
CompanyName2Vec	0.893	0.909	0.912
Fuzzy Distance Matching	0.852	0.857	0.865
Edit Distance Matching	0.565	0.589	0.614
Random Matching	0.001	0.002	0.004

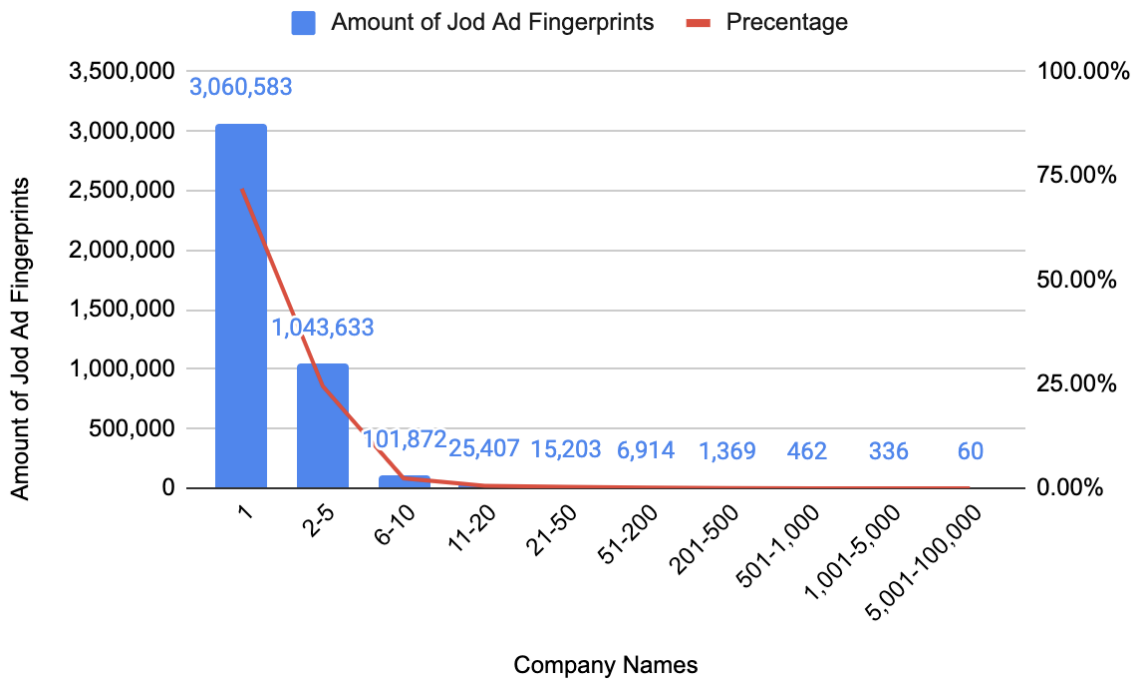


Figure 5.1: Fingerprints Distribution by Company Names

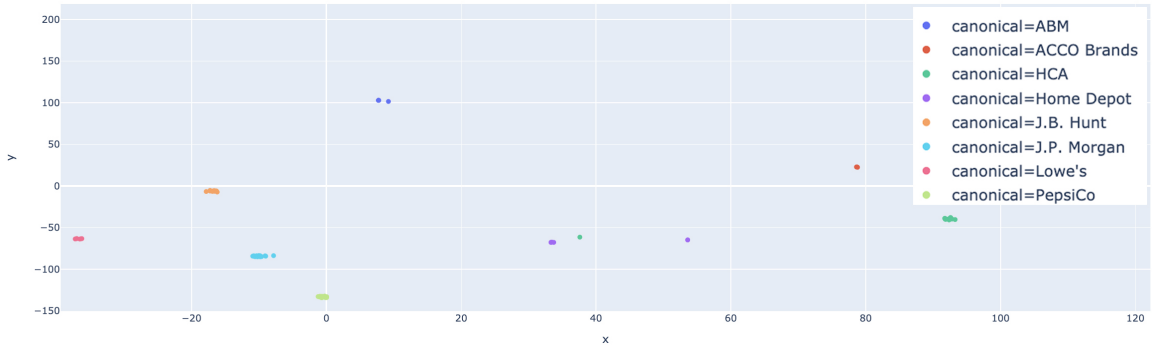
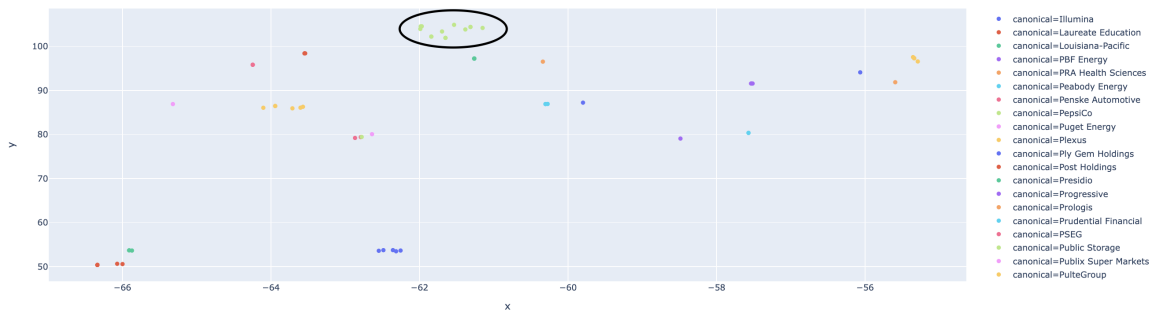
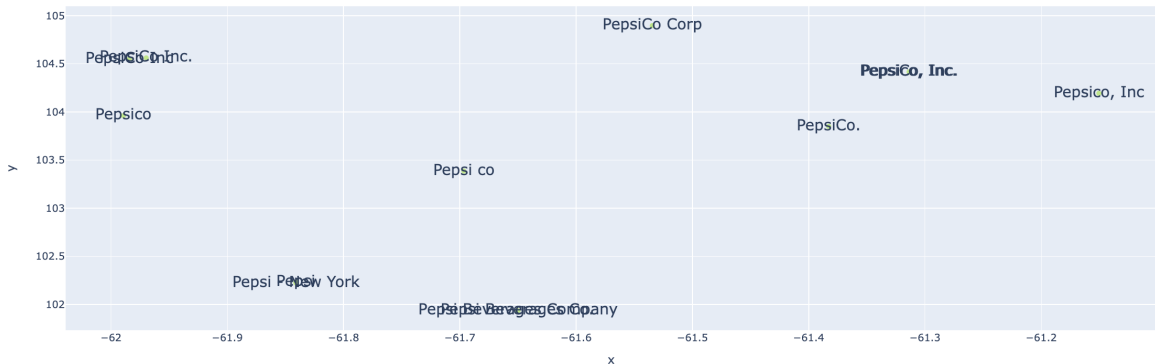


Figure 5.2: Synonym distribution of the following companies as were calculated by the CompanyName2Vec algorithm and reduced into two dimensions with t-SNE: ABM Industries, ACCO Brands, HCA Healthcare, Home Depot, J.B. Hunt Transport Services, Inc., JPMorgan Chase & Co., Lowe's Inc., and PepsiCo

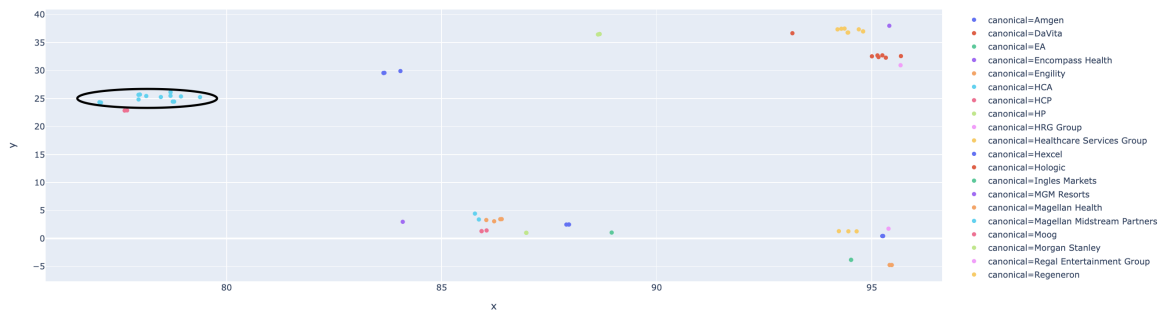


(a) Zoom out view on PepsiCo synonyms and other company synonyms near by

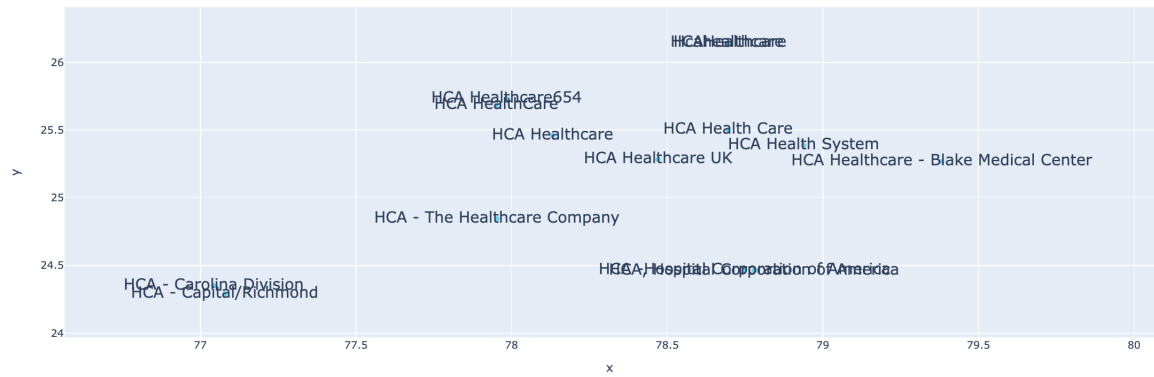


(b) Examples for PepsiCo synonyms

Figure 5.3: A plot of PepsiCo synonyms' vectors after were reduced into two dimensions

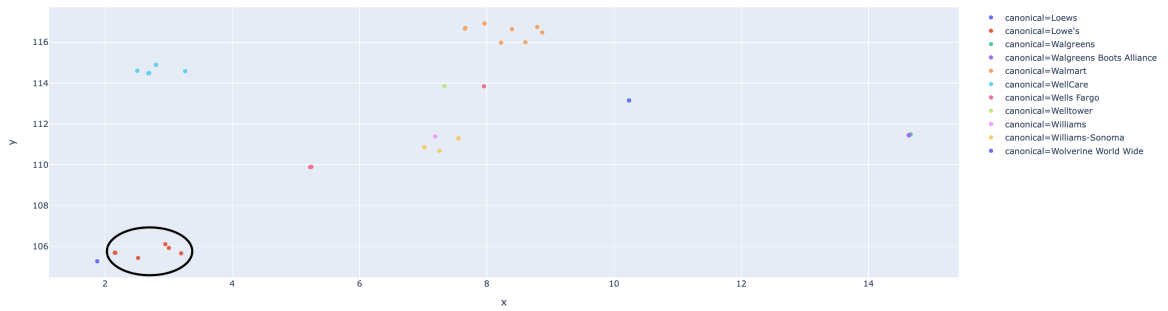


(a) Zoom out view on HCA Healthcare synonyms and other company synonyms near by

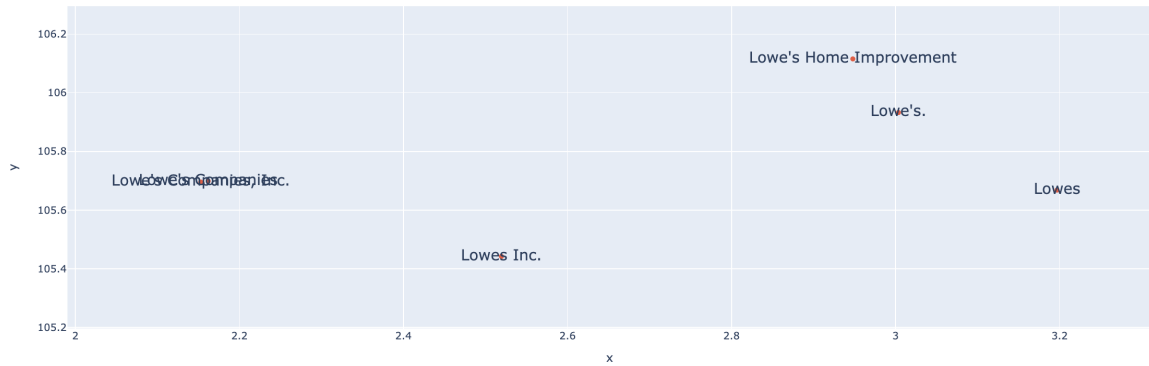


(b) Examples for HCA Healthcare synonyms

Figure 5.4: A plot of HCA Healthcare synonyms' vectors after were reduced into two dimensions

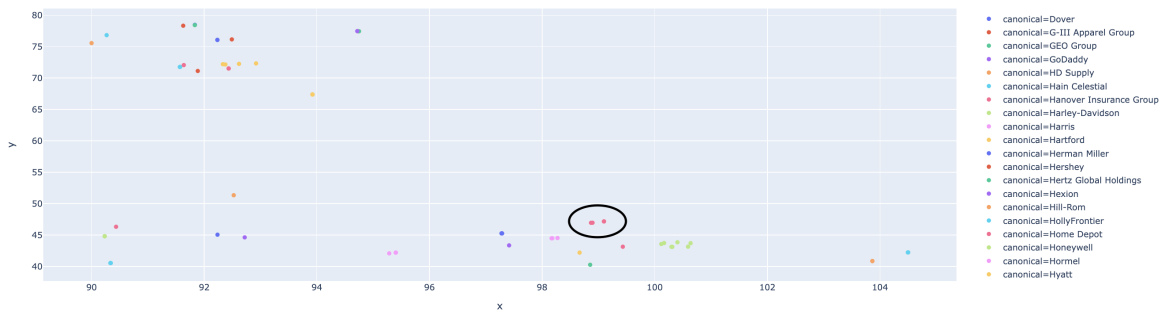


(a) Zoom out view on Lowe's Inc. synonyms and other company synonyms near by

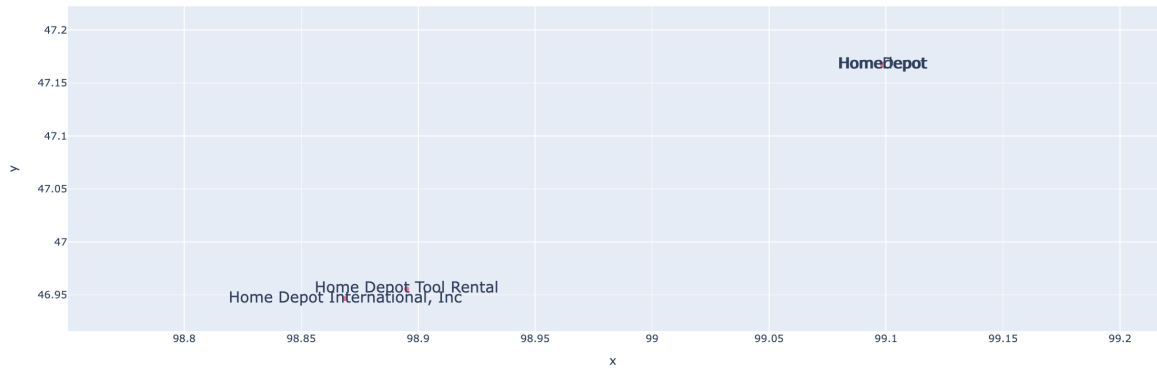


(b) Examples for Lowe's Inc. synonyms

Figure 5.5: A plot of Lowe's Inc. synonyms' vectors after were reduced into two dimensions

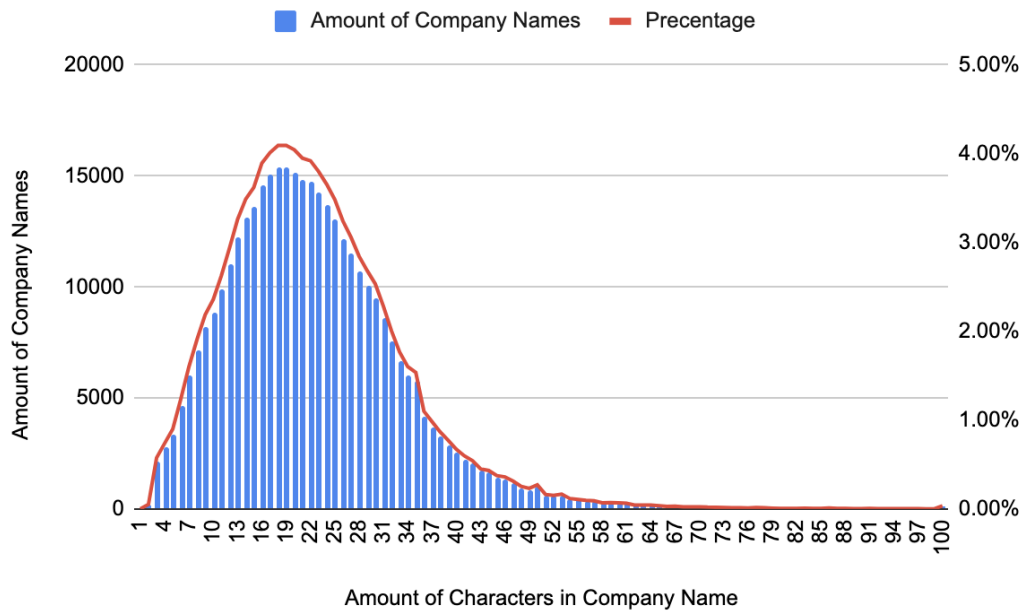


(a) Zoom out view on Home Depot synonyms and other company synonyms near by

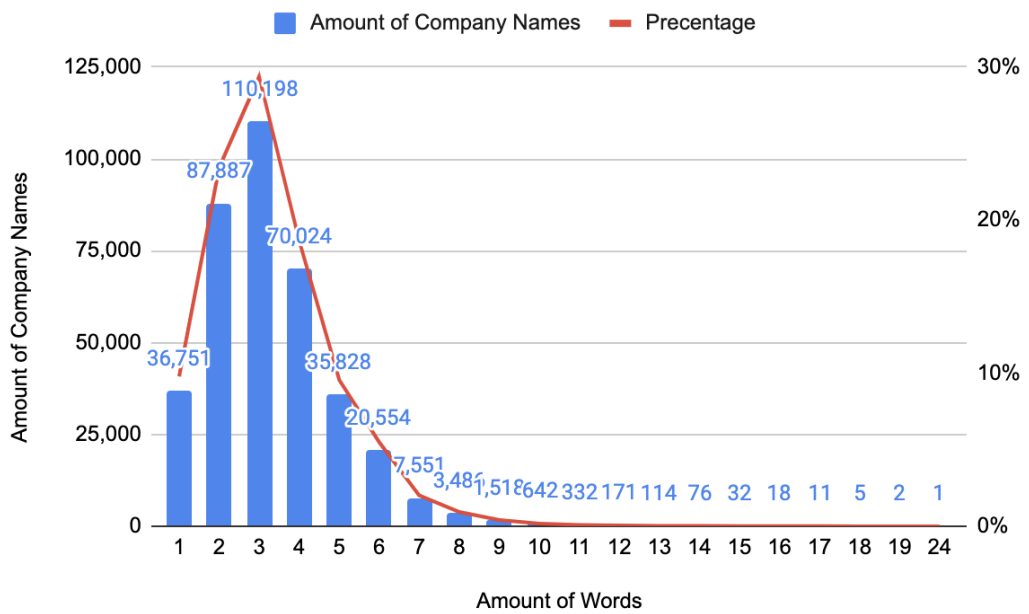


(b) Examples for Home Depot synonyms

Figure 5.6: A plot of Home Depot synonyms' vectors after were reduced into two dimensions



(a) Distribution of Company Names by Name's Length



(b) Distribution of company names by amount of words

Figure 5.7: Distribution of company names exist in job ad corpus

CHAPTER 6

DISCUSSION

Upon analyzing the results presented in the previous chapter, we can conclude the following: First, the novel approach we call `CompanyName2Vec`, which defines a vector representation of company names based on job ads, had promising results compared to other algorithms (see Table 5.1) and is helpful for the task of CEM (see Section 4.2). `CompanyName2Vec` algorithm captures the company names semantics and manages common prefixes, suffixes, and punctuation for company synonyms, like PepsiCo’s suffixes - Inc, Corp., Co., Company, Etc. (see Figure 5.3). `CompanyName2Vec` manages less popular suffixes as well, which may include a location or a point of interest. For example, HCA Capital Richmond is a synonym of HCA Healthcare. HCA Capital Richmond is a synonym due to the presence of HCA capital offices, which are located in Richmond, VA (see Figure 5.4). Another example is *Pepsi - New York*, which is a synonym of PepsiCo’s office in New York City (see Figure 5.3). A company’s business line can be attached to a company name as a synonym, such as Home Depot Tools Rental (see Figure 5.6), or Lowe’s Home Improvement (see Figure 5.5). Home Depot Tools Rental is a synonym for Home Depot, and Lowe’s Home Improvement is the company slogan.

Based on the method’s performance comparison (see Table 5.1), we found the `CompanyName2Vec` outperforms the other algorithms tested, while the second-best algorithm measured is the Fuzzy distance matching.

We found that “best partial” heuristic (see Subsection 2.2.1) works well in the case of matching company names because of its short length. The median length of a company name is 21 characters (see Figure 5.7b), and it consists in the median case of just three words (see Figure 5.7b).

We found the `CompanyName2Vec` method works better than the Fuzzy method mainly

because the `CompanyName2Vec` uses the language semantic. For example, trying to match “Pepsi - New York” with “PepsiCo Inc.” and “New York Life” using Fuzzy distance matching returns “New York Life” because New York is a significant part of both names. The `CompanyName2Vec` method returns “PepsiCo Inc.” based on an LSTM model, which extracts the company naming semantics from the training set.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this study, we introduced `CompanyName2Vec`, a novel, generic open-source algorithm which uses job ads and deep learning to address some of the challenges associated with company name synonyms. We provided a comprehensive description of our algorithm's steps, starting with a collection of job ads and finding company name synonyms using the job ads fingerprinting technique and generated company name synonyms datasets. We used text matching heuristics to reduce wrong synonyms and used the dataset to generate an Bi-LSTM model. This model generates a vector representation given any company name. We collected a set of synonyms for the Fortune 1,000 companies in the United States and used them to test and compare the `CompanyName2Vec` algorithm with several different matching algorithms. We matched all the Fortune 1,000 companies' synonyms with their canonical names and analyzed the results.

We make the following observations and conclusions:

1. The `CompanyName2Vec` algorithm was beneficial for confronting the problem of recommending synonyms for a given company name with higher $Success@k$ than other evaluated methods.
2. The `CompanyName2Vec` algorithm helps solve the CEM problem using job-ads data and requires no company's information for matching besides its name.

A possible future research direction is to examine additional configurations that may improve similar name suggestions:

- Using more comprehensive job ads' fingerprinting technique using additional information than just the job description, such as work location, contact information, company logo, Etc.

- Taking into account company abbreviations - due to lack of semantics in company name abbreviations, the CompanyName2Vec solution performs just as well as Fuzzy distance matching.
- Combining how the name sounds when spoken.
- Exploring other models than Bi-LSTM for name embedding, like bag of words, convolutional neural network (CNN), and bidirectional encoder representations from transformer (BERT).

Another avenue to pursue is learning company relations from the job ads corpus, for example, a parent company, a subsidiary, a holding company, Etc.

REFERENCES

- [1] Z. Sun, K. Strang, and S. Firmin, “Business analytics-based enterprise information systems,” *Journal of Computer Information Systems*, vol. 57, no. 2, pp. 169–178, 2017.
- [2] J. Liebowitz, “Business analytics and decision-making: The years ahead,” *The World Financial Review*, vol. 28, 2014.
- [3] Z. Sun, L. Sun, and K. Strang, “Big data analytics services for enhancing business intelligence,” *Journal of Computer Information Systems*, vol. 58, no. 2, pp. 162–169, 2018.
- [4] T. Gschwind, C. Miksovic, J. Minder, K. Mirylenka, and P. Scotton, “Fast record linkage for company entities,” in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 623–630.
- [5] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate record detection: A survey,” *IEEE Transactions on knowledge and data engineering*, vol. 19, no. 1, pp. 1–16, 2006.
- [6] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
- [7] D. Vesset, B. McDonough, D. Schubmehl, and M. Wardley, “Worldwide business analytics software 2013–2017 forecast and 2012 vendor shares (doc# 241689),” *Retrieved*, vol. 6, no. 28, p. 2014, 2013.
- [8] M. Paganelli, P. Sottovia, F. Guerra, and Y. Velegrakis, “Tuner: Fine tuning of rule-based entity matchers,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2945–2948.
- [9] L. Getoor and A. Machanavajjhala, “Entity resolution: Theory, practice & open challenges,” *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2018–2019, 2012.
- [10] I. Veldman, “Matching profiles from social network sites,” 2009.
- [11] S. Chaudhuri, V. Ganti, and R. Motwani, “Robust identification of fuzzy duplicates,” in *21st International Conference on Data Engineering (ICDE’05)*, IEEE, 2005, pp. 865–876.
- [12] O. Bousquet, U. von Luxburg, and G. Rätsch, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*. Springer, 2011, vol. 3176.

- [13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “Dbpedia: A nucleus for a web of open data,” in *The semantic web*, Springer, 2007, pp. 722–735.
- [14] K. K. Jones, S. N. Zenk, E. Tarlov, L. M. Powell, S. A. Matthews, and I. Horoi, “A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments,” *BMC research notes*, vol. 10, no. 1, p. 35, 2017.
- [15] L. Getoor and C. P. Diehl, “Link mining: A survey,” *Acm Sigkdd Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.
- [16] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, “Swoosh: A generic approach to entity resolution,” *The VLDB Journal*, vol. 18, no. 1, pp. 255–276, 2009.
- [17] D. G. Brizan and A. U. Tansel, “A. survey of entity resolution and record linkage methodologies,” *Communications of the IIMA*, vol. 6, no. 3, p. 5, 2006.
- [18] I. P. Fellegi and A. B. Sunter, “A theory for record linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.
- [19] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
- [20] G. Navarro, “A guided tour to approximate string matching,” *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.
- [21] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.
- [22] W. E. Yancey, “Evaluating string comparator performance for record linkage,” *Statistics*, vol. 5, p. 38, 2005.
- [23] S. Zhang, Y. Hu, and G. Bian, “Research on string similarity algorithm based on levenshtein distance,” in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, IEEE, 2017, pp. 2247–2251.
- [24] SeatGeek, *Fuzzy-wuzzy python package*, (accessed July 16, 2021).
- [25] W. W. Cohen, “Integration of heterogeneous databases without common domains using queries based on textual similarity,” in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 201–212.

- [26] D. Holmes and M. C. McCabe, “Improving precision and recall for soundex retrieval,” in *Proceedings. International Conference on Information Technology: Coding and Computing*, IEEE, 2002, pp. 22–26.
- [27] A. J. Lait and B. Randell, “An assessment of name matching algorithms,” *Technical Report Series-University of Newcastle Upon Tyne Computing Science*, 1996.
- [28] P. Christen, “A comparison of personal name matching: Techniques and practical issues,” in *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW’06)*, IEEE, 2006, pp. 290–294.
- [29] N. Elbegbayan *et al.*, “Winnowing, a document fingerprinting algorithm,” *TDDC03 Projects, Spring*, 2005.
- [30] S. Schleimer, D. S. Wilkerson, and A. Aiken, “Winnowing: Local algorithms for document fingerprinting,” in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 76–85.
- [31] J. Parapar and Á. Barreiro, “Winnowing-based text clustering,” in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 1353–1354.
- [32] A. T. Wibowo, K. W. Sudarmadi, and A. M. Barmawi, “Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on bahasa indonesia documents,” in *2013 International Conference of Information and Communication Technology (ICoICT)*, IEEE, 2013, pp. 128–133.
- [33] A. Aiken, “Moss: A system for detecting software similarity,” vol. 29, p. 2017, 1994, (accessed July 12, 2021).
- [34] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” in *Advances in Neural Information Processing Systems*, 2001, pp. 932–938.
- [35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [37] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [38] A. Frome *et al.*, “Devise: A deep visual-semantic embedding model,” 2013.

- [39] P. Li, Y. Liu, M. Sun, T. Izuba, and D. Zhang, “A neural reordering model for phrase-based translation,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1897–1907.
- [40] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, “Bilingually-constrained phrase embeddings for machine translation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 111–121.
- [41] J. Zhao, M. Lan, Z.-Y. Niu, and Y. Lu, “Integrating word embeddings and traditional nlp features to measure textual entailment and semantic relatedness of sentence pairs,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2015, pp. 1–7.
- [42] B. Salehi, P. Cook, and T. Baldwin, “A word embedding approach to predicting the compositionality of multiword expressions,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 977–983.
- [43] E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin, “Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning,” *arXiv preprint arXiv:1509.01692*, 2015.
- [44] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.
- [45] R. Rivest and S. Dusse, *The md5 message-digest algorithm*, 1992.
- [46] K. Greff, R. K. Srivastava, J. Koutnk, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [47] A. F. Agarap, “Deep learning using rectified linear units (relu),” *arXiv preprint arXiv:1803.08375*, 2018.
- [48] U. B. of Labor Statistics, *Job openings and labor turnover summary*, (accessed June 21, 2021).
- [49] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

Appendices

APPENDIX A
REFERENCES TO PROJECT CODE AND DATASETS

Code and datasets can be found at the following link:

<https://tinyurl.com/he4tzk>

תקציר

התאמת ישויות (Entity Matching) היא חלק חיוני בכל המערכות בעולם האמיתי המקבלות נתונים מובנים (Structure Data) ולא מובנים (Un-Structure Data) המגיעים ממקורות שונים. בדרך כלל אין מפתח משותף זמין לחיבור רשומות. תהליכי ניקוי ואינטגרציה מאסיביים של הנתונים נדרשים לצורך השלמת מיידעים והעשרת המידע לפני שניתן לבצע ניתוח נתונים או עיבוד נוסף. למרות שההצמדה של רשומות נחשבת לעתים קרובות כצעד קצת מייגע אך הכרחי, היא חושפת תובנות יקרות ערך, תומכת ביכולת הדמיה וניתוח מתקדם של הנתונים ומהווה בסיס לגישות אנליטיות נוספות בניתוח נתונים (Data Analytics). בעבודה זו, אנו מתמקדים בהתאמת ישויות ארגוניות. אנו מציגים את CompanyName2Vec, אלגוריתם חדש לפתרון Company Entity Matching – CEM, באמצעות מודל רשת עצבית כדי ללמוד סמנטיקה של שם חברה באמצעות מודעות דרושים, מבלי להסתמך על מידע כלשהו על החברה המתאימה מלבד שמה. בהתבסס על מידע מתויג על שמות חברות מהעולם האמיתי, אנו מראים כי CompanyName2Vec עולה על שיטות מוערכות אחרות עם דיוק ממוצע של 0.894 לתוצאה יחידה.

עבודה זו בוצעה בהנחייתו של דר' מיכאל פייר מהמחלקה להנדסת מערכות תוכנה ומידע, הפקולטה
למדעי ההנדסה, אוניברסיטת בן גוריון, ובליוי דר' אילן גרונאו מבית הספר אפי ארזי למדעי המחשב,
אוניברסיטת רייכמן.



אוניברסיטת רייכמן
בית-ספר אפי ארזי למדעי המחשב
התכנית לתואר שני (M.Sc.) - מסלול מחקרי

CompanyName2Vec: Company EntityMatching Based on Job Ads

מאת
רן זיו

עבודת תזה המוגשת כחלק מהדרישות לשם קבלת תואר מוסמך M.Sc.
במסלול המחקרי בבית ספר אפי ארזי למדעי המחשב, אוניברסיטת רייכמן

ספטמבר 2021